Genetic prediction and adverse selection*

Eduardo Azevedo† Jonathan Beauchamp‡ Richard Karlsson Linnér§

October 28, 2024

Abstract

The predictive power of genetic data has been increasing rapidly and is reaching levels of clinical utility for many diseases. Meanwhile, many jurisdictions have banned insurers from utilizing genetic information. This has led to concerns that further improvements in genetic prediction will lead to adverse selection. We make three contributions to this debate. First, we develop a method to measure the amount of selection in an insurance market where consumers have access to current genetic prediction technology. Second, we extend the method to estimate the amount of selection given expected improvements in genetic prediction technology. Third, using the UK Biobank dataset with nearly 500,000 genotyped individuals, we apply the method to the critical illness insurance market. We find that expected improvements in genetic prediction are likely to lead to unsustainably high levels of selection and thus threaten the viability of the market. We discuss policy implications.

Keywords: adverse selection, asymmetric information, complex diseases, genetic risk prediction, genetic testing, insurance, polygenic indexes

JEL classifications: D82, G22, G28, I13, I18

*For helpful comments and suggestions, we thank Ralph Koijen, Martin Henriksson, Patrick Turley, as well as conference and seminar participants at Stanford University, Uppsala University, Washington University in St-Louis, the Yale School of Management, the American Enterprise Institute, The Advances in Social Genomics Conference, the Integrating Genetics in the Social Sciences conference, and the NBER Insurance Working Group meeting. For outstanding research assistance, we are grateful to Gareth Markel. This research has been conducted using the UK Biobank Resource under application numbers 11425, 99086, and 217716. This research was supported by the Dutch Research Council (NWO; project number Vl.Veni.221E.080), by a Small Compute grant (NWO/SURF, no. EINF-8891), by a Faculty Research and Development Award (FRDA) at George Mason University, and by the National Science Foundation (NSF award nos. 2343735 and 2343736). For extended acknowledgments, see the Online Appendix.

†The Wharton School, University of Pennsylvania. Email: eazevedo@wharton.upenn.edu

‡Interdisciplinary Center for Economic Science and Department of Economics, George Mason University. Email: jonathan.pierre.beauchamp@gmail.com

§Department of Economics, Leiden University: Email: r.karlsson.linner@law.leidenuniv.nl
School of Business and Economics, Vrije Universiteit Amsterdam: Email: r.karlssonlinner@vu.nl

1 Introduction

Over the past decades, technological and scientific advances have led to substantial increases in the predictive power of genetic data (Abdellaoui et al., 2023). Scientists can now construct **"polygenic indexes"** (**PGIs**) from a person's genetic data that partially predict their risk of developing various common diseases, such as breast or prostate cancer. The predictive accuracy of these PGIs recently reached levels of clinical and personal utility (Khera et al., 2018) and is expected to continue to improve and to approach its theoretical upper limits in the coming decades (Torkamani et al., 2018; Visscher et al., 2021; Abdellaoui et al., 2023). PGIs for common diseases are complementary to the already established use of diagnostic genetic tests for rare single-gene disorders (e.g., to diagnose Tay-Sachs disease), and clinical trials are underway to evaluate the integration of PGIs in routine healthcare and screening procedures (Klarin and Natarajan, 2022; Saya et al., 2022).

Tens of millions of consumers worldwide have already bought genetic tests, primarily to explore their genealogy (Regalado, 2019). The major consumer genetic companies, such as 23andMe and MyHeritage, have started to bundle their consumer genetic testing services together with PGI-based disease risk assessments (23andMe, 2023; MyHeritage, 2021). The price of these assessments is now often less than $150—similar to that of at-home blood or cholesterol tests—and is rapidly declining. Large healthcare providers, such as the UK's National Health Service (NHS), are planning to inform millions of patients about their PGIs in the next few years (Government Office for Science, 2022; Our Future Health, 2022). Because of these developments, many experts predict that both patients and consumers will be routinely informed of their PGI-based disease risk prediction in the near future (Zeggini et al., 2019; Meyer et al., 2024).

Meanwhile, in response to concerns that individuals may be treated inequitably on the basis of their genetic data, many jurisdictions have banned genetic discrimination in various sectors (Golinghorst et al., 2022).[1] Such bans typically prohibit insurers from requesting genetic information known to applicants (e.g., from a consumer genetic test) or from utilizing any genetic information at all—creating a distinct information asymmetry. Scholars, policymakers, and insurance industry stakeholders have voiced concerns that these developments

[1]For instance, in the US, the Genetic Information Nondiscrimination Act (GINA) of 2008 prohibits the use of genetic information by health insurers. Some states have also legislated state-level bans to limit the use of genetic information by other types of insurers. In Canada, the Genetic Non-Discrimination Act (GNDA) of 2017 broadly prohibits genetic discrimination, including in all sectors of the insurance industry. Similarly, in France, the Penal Code prohibits the use of genetic information other than for medico-scientific purposes, and thus by all insurers (Bélisle-Pipon et al., 2019). Some countries—including the UK, the Netherlands, and Sweden—generally prohibit the use of genetic information but have exceptions for especially large insurance coverage amounts or specific disorders such as Huntington's disease (Nabholz and Rechfeld, 2017).

will lead to high levels of adverse selection in some segments of the insurance industry (Joly et al., 2014; Rechfeld et al., 2019; Maxwell et al., 2021). This could in principle lead to spiraling insurance premiums and to the unraveling of important market segments, and thus leave swathes of consumers—including the most vulnerable ones—uninsured (Akerlof, 1970; Einav et al., 2023). However, despite decades of research on the potential impact of genetic testing on insurance markets, there is still considerable debate over whether PGIs actually pose a major threat (Golinghorst et al., 2022; Yanes et al., 2024; Dixon et al., 2024).

This study combines large-scale genetic data, economic theory, and econometrics to inform the debate on genetics and adverse selection in insurance markets. We focus on the impact of PGIs on the market for **critical illness insurance** (**CII**), which is estimated to be worth more than $100 billion globally and is growing rapidly (Allied Market Research, 2022). CII pays out a lump sum in the event that the insured person gets diagnosed with any of a number of medical conditions listed on the policy (Brackenridge et al., 2006). [2] We focus on the CII market because it is an ideal testbed for the methodology we develop.

This study makes three key contributions. First, we develop a method to measure the amount of adverse selection that would take place if consumers had widespread access to current genetic information. The method uses large-scale data on genetics, non-genetic risk factors, and insured events. Second, we extend the method to estimate the amount of selection that will occur with future genetic prediction technology. The idea is to combine estimates from heritability studies[3]—which are informative of the likely accuracy of future PGIs—with estimates of current PGIs' correlations with non-genetic covariates that are commonly used by insurers in underwriting. The method's structural assumptions are grounded in empirical regularities from genetics.

Third, we apply the method to measure selection in the CII market with both current and future genetic prediction technology. For this, we use data from the UK Biobank (UKB) (Bycroft et al., 2018), a large dataset of 500,000 individuals with genotypic and rich health-related data, including linked data from the UK's National Health Service (NHS). To illustrate our method in a simple setting, we first consider seven single-disease CII contracts—for Alzheimer's disease, breast cancer (for females), coronary artery disease, colorectal cancer, prostate cancer (for males), schizophrenia, and type 2 diabetes—that each pay a lump sum upon the onset of the disease. We next apply our method to multiple-disease contracts that bundle the diseases and pay a lump sum upon the onset of any of the diseases (as in the most common CII contracts).

[2] Commonly covered conditions include different types of cancers, coronary artery disease, stroke, multiple sclerosis, paralysis, blindness, and unspecified terminal illness (Gatzert and Maegebier, 2015).

[3] The heritability of a trait or disease is the share of the variation in the trait or disease that is attributable to genetic factors.

We report three main findings. First, with current genetic prediction technology, we find noticeable levels of selection. For Alzheimer's disease, prostate cancer, and schizophrenia, we estimate selection levels that are similar to those found in previously studied market segments that had unraveled. However, these results are based on the assumption that genetic prediction technology is widespread, which is not yet the case. If we instead more realistically assume incomplete take-up of the current genetic prediction technology, selection becomes limited for all single-disease contracts. Further, for the multiple-disease contract, we do not find noticeable amounts of selection. These results are consistent with the fact that life and critical illness insurance markets do not yet appear to be plagued by high levels of selection from genetic tests (Golinghorst et al., 2022).

Second, with the expected future genetic prediction technology, the degree of selection becomes extremely high. For all contracts, we find degrees of selection that are within or above the range observed in market segments that had unraveled. Even in a scenario where only 50% of consumers privately perform genetic tests, we observe high levels of selection. These results suggest that the commonplace policy (in developed countries) of simply banning insurers from using genetic data may become problematic when genetic prediction improves. Some market segments may collapse, leaving consumers without insurance. Other, possibly complementary, policy tools should be considered.

Third, we find that there is subtle variation in the amount of selection for different contracts. The amount of selection is mostly determined by the expected accuracy of future prediction technology, by how well observable non-genetic risk factors can predict the diseases, and by the prevalence of the diseases.

A key driver of these findings is that we consider genetic predictions based on PGIs, which aggregate information from over one million genetic variants to predict a substantial share of the variation in disease risk. For instance, the current PGI of prostate cancer accounts for 11.2% of the variation and—based on heritability studies—we expect future PGIs to account for 18 to 58% of the variation. By contrast, popular discussion of genetic prediction often focuses on rare mutations that individually have large effects on the risk of contracting some diseases. Well-known examples include mutations in the *BRCA* genes for breast cancer and in the *PKD* genes for polycystic kidney disease. Interestingly, these mutations are unlikely to generate much adverse selection because they are rare and can often be predicted from family history.

Our findings complement a literature that finds that consumers' insurance choices are responsive to genetic information. Oster et al. (2010) find that carriers of the genetic variant that causes Huntington's disease (a rare single-gene disorder affecting less than 0.01% of the population) were 5 times more likely to own long-term care insurance compared to the

general population. Taylor Jr et al. (2010) find that carriers of the *APOE4* variant, which increases the risk of Alzheimer's by a factor of 2 to 3, were 2.3 times more likely to buy (or plan to buy) long-term care insurance relative to carriers of the more common *APOE3* variant.

Existing research on the potential impact of genetic testing on insurance markets has focused almost exclusively on tests for rare genetic mutations and has obtained mixed results. Macdonald and McIvor (2009) and Macdonald (2011) find little impact, while Howard (2016) shows that the impact could be substantial. Our study is the first (to our knowledge) to tackle this question with large-scale genetic and health data together with an econometric model that accounts for the predictive power of future PGIs over and above that of observable risk factors.

Given our findings of considerable potential for selection, we discuss policy remedies at the end of the paper. Our findings do not imply that the laissez-faire approach of letting insurers use genetic information as they wish without governmental intervention is necessarily optimal. Other policy tools, such as community rating, subsidies, and risk adjustment, may strike a better balance between the goals of having well-functioning insurance markets and maintaining access for vulnerable consumers.

1.1 Illustrative example: CII for prostate cancer

It is possible to understand the gist of our contribution with a simple example. Consider our single-disease contract for prostate cancer. This one-time contract pays $100,000 if the buyer ever incurs prostate cancer by age 65. This contract is a simplification of real CII contracts but, as we will discuss later, it captures several key features. The contract is purchased at age 35, many years before the usual age of onset of prostate cancer. The insurer's cost of selling this contract is roughly proportional to the buyer's risk of contracting prostate cancer by age 65. Our methodology allows us to calculate the distribution of risk in a population conditional on various subsets of information. We will focus on information subsets that comprise observable covariates used by insurers in underwriting, as well as on combinations of these covariates with a current or future PGI.

The left panel of Figure 1 shows results based on UKB data. Scenario 1 shows the distribution of the risk of contracting prostate cancer by age 65 in an all-male population of typical consumers, conditional on non-genetic covariates only. This information is observed by both consumers and insurers and is used by insurers to set prices. There would be a non-trivial amount of selection if insurers were not allowed to use this information. But in practice, insurers do underwrite and define various consumer risk classes which they price

differently. Following standard practice in the insurance industry, we define the **standard risk class** as the set of consumers whose risk of contracting prostate cancer is between 0.75 and 1.25 times the average population risk ($\bar{r} = 0.053$) (Brackenridge et al., 2006). These consumers are shown in blue in the figure and face the same price of insurance. By definition, these consumers have almost the same risk conditional on the non-genetic covariates, and so there is almost no selection within the standard risk class in Scenario 1.

Scenario 2 shows the distribution of risk conditional on both the non-genetic covariates and the current PGI of prostate cancer. The risk distribution for the standard risk class now has a wider spread. Because insurers are not allowed to observe genetic information, these consumers look like nearly identical risks to insurers and are therefore charged the same price. However, because consumers can observe their PGIs, their privately known risk varies widely.

Scenarios 3L and 3U show the distribution of risk with two future PGIs with different assumed accuracies. These assumed accuracies are informed by existing estimates of the heritability of prostate cancer and correspond to what we consider reasonable lower (Scenarios 3L) and upper (Scenarios 3U) bounds. In Scenario 3L, the standard risk class now has an even wider spread, with consumers at the 5th and 95th percentiles facing risks of 0.4% and 15.9% of contracting prostate cancer, respectively. And in Scenario 3U, consumers at the 5th percentile face a virtually null risk (as can be seen from the spike at 0) vs. a 30.1% risk for consumers at the 95th percentile. Thus, consumers who face the same price of insurance vary widely with respect to their privately known risk. This suggests that adverse selection could be a serious issue.

To quantify the degree of asymmetric information and adverse selection that would result from this situation, we employ the **implicit tax**, a metric proposed by Hendren (2013, 2017). The right panel of Figure 1 shows for each scenario the estimated implicit tax as a function of a consumer's risk percentile for prostate cancer, for the consumers in the standard risk class. In Scenarios 2, 3L, and 3U, the implicit taxes imply very high levels of selection and exceed those that have been estimated for insurance market segments that have unraveled. This, along with the results we present below for the other CII contracts we study, suggests selection due to future genetic prediction technology will be a first-order concern and may make some insurance market segments non-viable.

1.2 Paper structure

The remainder of the paper is structured as follows. Section 2 provides background information on genetic prediction and on the CII market. Section 3 describes our theory, introducing

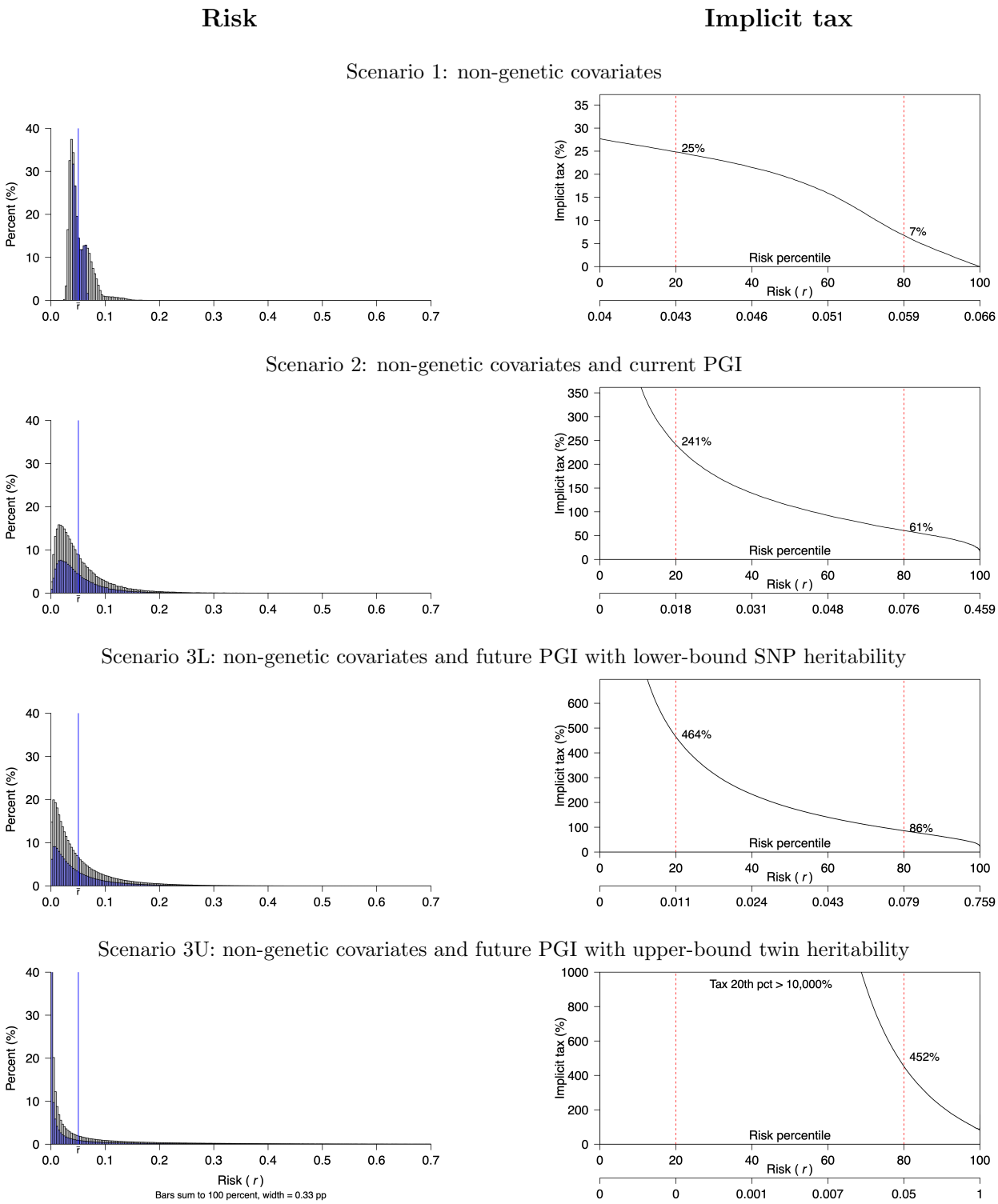

Figure 1: Single-disease CII contract for prostate cancer: risk and implicit tax

Notes: Left panel: distribution of the risk of contracting prostate cancer by age 65, conditional on various information scenarios and computed in the UKB data ($N = 181,627$). The vertical blue line marks the average risk in the standard risk class ($\bar{r} = 0.05$). The standard risk class individuals are shown in blue; insurers treat these individuals identically, so the blue distribution corresponds to the distribution of private risk for these individuals. Right panel: implicit tax for consumers in the standard risk class as a function of their percentile private risk of prostate cancer, for each scenario.

our economic and econometric models. Section 4 describes our data and our empirical specification. Section 5 presents our results, first focusing on the prostate cancer single-disease contract, then on all the single-disease contracts, and finally on the multiple-disease contracts. Section 6 discusses the robustness of our results and some extensions, while Section 7 discusses policy implications. Section 8 concludes.

2 Background

2.1 Genetics background

Some disorders, such as Huntington's disease and sickle cell anemia, are caused by single genetic variants. These disorders are called monogenic. Some traits, like eye color, are principally influenced by a relatively small number of variants (Simcoe et al., 2021). However, the vast majority of traits are polygenic: they are influenced by a very large number of causal genetic variants (as well as by environmental factors), with each causal genetic variant having a tiny effect (Chabris et al., 2015; Holland et al., 2020; O'Connor et al., 2019). For example, coronary artery disease, colorectal cancer, prostate cancer, schizophrenia, and type 2 diabetes are polygenic diseases, and height, BMI, and cognitive ability are polygenic continuous traits.[4] We focus on polygenic diseases.

Consider first how genetic data are typically used to predict a continuous polygenic trait Y, such as height. Suppose we have data on a set of genetic variants from across the genome. As we explain in more detail in Online Appendix A, the entire genome contains more than 3 billion pairs of elements (associated with the letters "A", "C", "G", and "T"), but for most traits, genotyping data on $\sim$1 million variants capture an important share of the relevant genetic variation.[5]

Y is assumed to follow the **additive genetic model** (Falconer and Mackay, 1996), which posits that there are no interactions among variants or between variants and the environment, and that genetic effects are linear with respect to the variants. Accordingly, Y is the weighted sum of an individual's genetic variants, where the weights are the variants' true effect sizes on Y, plus a disturbance term ψ that captures all other influences:

$$Y = k + \mathbf{X}\,\mathbf{b} + \psi,$$

[4] Alzheimer's disease and breast cancer are also polygenic, but in addition to being affected by a large number of variants with tiny effects, they are affected by one common variant (in the *APOE* gene, for Alzheimer's) or by rare variants (in the *BRCA* genes, for breast cancer) with large effects. Still, the bulk of the genetic variation for Alzheimer's and breast cancer is due to many variants with small effects.

[5] Genotyping is the process of identifying a subset of an individual's DNA that tags much of the genome's variation; sequencing is the process of identifying every single element of an individual's DNA (and is costlier).

where k is a constant, $\mathbf{X}$ is a row vector that contains the measured variants, and $\mathbf{b}$ is a column vector that contains their true effect sizes on Y. $G^* = \mathbf{X}\,\mathbf{b}$ is the individual's **true additive genetic factor** for Y. (Throughout, we denote random variables with uppercase and their realizations and constants with lowercase letters, and vectors with bold font.) Despite its simplicity, the additive genetic model is commonly used in genetics, and both theory and empirical evidence suggest that additive effects account for the bulk of the genetic variation for most polygenic diseases (Hill et al., 2008; Polderman et al., 2015).

In practice, neither $\mathbf{b}$ nor G^* are observed. Instead, one can obtain estimates $\hat{\mathbf{b}}$ from a previously published **genome-wide association study** (**GWAS**)[6] of the trait and use them to construct a **polygenic index** (**PGI**) of Y:

$$G = \mathbf{X}\,\hat{\mathbf{b}}.$$

Thus, the observed index G is a noisy measure of the true additive genetic factor G^*.[7] Let ϵ be the error: $G = G^* + \epsilon$. ϵ stems from measurement error in $\hat{\mathbf{b}}$, which is due to sampling variation in the GWAS. To avoid overfitting, the GWAS is conducted in a sample that is independent of the analysis dataset in which the PGI is constructed, so ϵ and G^* are independent. Online Appendix A provides further background on the genetics of polygenic traits, on GWASs, and on PGIs.

As mentioned, a polygenic trait is influenced by a large number of genetic variants, each with a tiny effect. This implies, by the central limit theorem, that $G^* = \mathbf{X}\,\mathbf{b}$ is approximately normally distributed. By the same reasoning, so is ϵ. As a result, so is G—i.e., PGIs also are approximately normally distributed. This can be seen in Online Appendix Figure A.2, which shows the distributions of the PGIs we construct for the diseases we study.[8]

This model for a continuous trait can easily be adapted to a **disease** D. Following standard practice in epidemiological genetics, we model D with a liability threshold model, according to which one contracts the disease if one's **liability** $\mathcal{L}$ is positive. The above model still applies, but with $\mathcal{L}$ substituted for Y. ψ is often assumed to be normally distributed, in which case the liability threshold model becomes a probit regression of the disease. As

[6] A GWAS is a large-scale genetic study in which a trait of interest is regressed on each of millions of genetic variants, separately. As we explain in Online Appendix A, the method we use to construct the PGIs, PRS-CS (Ge et al., 2019), computes the estimates $\hat{\mathbf{b}}$ by adjusting the GWAS regression estimates to account for the effects of nearby, correlated SNPs.

[7] The GWAS and $\hat{\mathbf{b}}$ estimates are measures of the variants' *associations* with the trait, rather than of their effects on the trait (below, for simplicity, we at times nonetheless refer to $\hat{\mathbf{b}}$ as effects). The PGIs we construct do not therefore have a clear causal interpretation. They are, however, useful for prediction, which is what we need here.

[8] Our PGI of Alzheimer's disease is not normally distributed because it strongly tags the large-effect *APOE* variants.

Online Appendix Figure A.3 shows, for all the diseases we study, predicted disease risk under the liability threshold model closely matches observed risk.

As GWASs have rapidly grown in sample size over the past decade (Loos, 2020), estimates of $\mathbf{b}$ have become much more precise; as a result, the predictive power of PGIs has also increased dramatically (see, e.g., Okbay et al., 2022). As GWAS sample sizes keep increasing, and as the set of genetic variants they analyze is augmented to include rarer variants, PGI predictive power should increase further (Abdellaoui et al., 2023; Daetwyler et al., 2008). In theory, the upper bound for a PGI's predictive power for a given trait is the trait's narrow sense heritability (So and Sham, 2010; Dudbridge, 2013), defined as the share of the trait variation that is attributable to additive genetic factors.

Table 1 shows, for the seven diseases we study, our estimates of their current PGIs' R^2's together with previous studies' estimates of their "SNP heritabilities" and of their "twin heritabilities". Since the diseases are binary traits, we report their R^2 **on the liability scale**, also known as the McKelvey & Zavoina pseudo-R^2. This R^2 is defined as $\text{Var}[\lambda G]/\text{Var}[\mathcal{L}_r]$, where λ is the coefficient from a probit regression of D on G and $\text{Var}[\mathcal{L}_r] = \text{Var}[\lambda G]+1$ is the variance of the liability in that regression (Lee et al., 2012; McKelvey and Zavoina, 1975).[9] A disease's **SNP heritability** is the share of the variance in its liability $\mathcal{L}$ that is attributable to the set of common genetic variants used in the SNP heritability estimation.[10] Because not all variants are used (as most were not measured, and for computational reasons), SNP heritability is necessarily lower than the true heritability. A trait's **twin heritability** is computed by comparing the resemblance of monozygotic twins—who share 100% of their genomes—to that of dizygotic twins—who share 50% on average. Under some assumptions, it is a consistent estimate of the trait's true heritability.

While it is impossible to know with certainty how predictive a disease's future PGI will be, it is likely that it will explain at least as much as the disease's SNP heritability estimated with common variants. This is because future PGIs, in addition to being constructed with more precise estimates of common variants' effect sizes, will also account for the effects sizes of less common variants. And, as mentioned, the theoretical upper limit on a PGI's predictive power for a disease is the disease's heritability, which we estimate with the twin heritability. In our analyses, we report results under the assumptions both that future PGIs' accuracies will be equal to their SNP heritabilities (as lower bounds) and their twin heritabilities (as

[9]The R^2 on the liability scale is intuitively analogous to the usual R^2, but indicates the share of the variance of the liability—instead of the variance of a continuous dependent variable—that is accounted for by the covariates.

[10]The expression "SNP heritability" comes from the fact that the variants used are typically single nucleotide polymorphisms (SNPs). The SNP heritability estimates in Table 1 come from different sources (listed in Supplementary Table 1) and were obtained using different estimation pipelines and sets of common SNPs. They must be seen as ballpark figures rather than precise point estimates.

Table 1: Descriptive statistics for selected diseases

Disease	**Prevalence**	**Lifetime risk**	**Current PGI R^2**	**SNP h^2**	**Twin h^2**
Alzheimer's disease	7.1%	6.3% (M), 12% (F)	6.5%	31%	58%
Breast cancer (F only)	8.1%	12.3%	6.8%	14%	31%
Coronary artery disease	3.0%	51.7% (M), 39.2% (F)	3.0%	22%	40%
Colorectal cancer	2.9%	4.6%	2.5%	9%	40.2%
Prostate cancer (M only)	0.9%	12.7%	11.2%	18%	57%
Schizophrenia	0.46%	0.7%	5.8%	24%	79%
Type 2 diabetes	3.9%	27% (M), 18.6% (F)	8.6%	19.6%	72%

Notes: Current PGI R^2's are on the liability scale and were estimated in the UKB. All other statistics come from earlier, separate studies of the diseases and were not produced with a uniform methodology across diseases. References for these studies are listed in Supplementary Table 1. Prevalence and lifetime risk include various age-corrected measures (e.g., the prevalence of Alzheimer's disease is reported for ages above 65).

upper bounds).

As is currently standard practice in applied genetic research, our analysis only uses data from individuals of European ancestry, for two main reasons. First, available genetic data for other ancestries are still scarce (Mills and Rahal, 2020); as a result, most large-scale GWASs whose estimates we can use to construct our PGIs were conducted among European ancestry individuals. Second, PGIs based on European ancestry GWAS estimates tend to perform poorly among individuals of other ancestries (Martin et al., 2017; Privé et al., 2022).[11] There are numerous ongoing efforts to diversify the world's biobanks (Zhou et al., 2022) and to develop statistical methods to improve polygenic prediction across ancestries. As a result, PGI predictive power for non-European ancestries will also increase dramatically in the future. Our main qualitative results regarding the implications of improved genetic prediction for insurance markets are thus also relevant for non-European ancestry individuals.

2.2 Critical illness insurance

Critical illness insurance (**CII**) pays out a lump sum in the event that the insured person gets diagnosed with any of the medical conditions listed on the policy (Brackenridge et al., 2006). The lump sum can be used as the policyholder wishes. The policy pays out once and is thereafter terminated. The typical policy is renewable term at guaranteed rates, which are determined before contract initiation by medical underwriting.

Typical contracts cover some 20 to 50 core illnesses or medical conditions, including different types of cancers (e.g., breast, prostate, colorectal, and lung cancer), coronary artery disease, multiple sclerosis, stroke, as well as serious conditions like paralysis, blindness, and unspecified terminal illness (Gatzert and Maegebier, 2015). Enhanced plans exist that cover

[11] This is likely due to population genetic reasons and differential genetic effects in different environments.

additional conditions, such as dementia, diabetes, or severe psychiatric illness. Though CII contracts typically cover many conditions, cancers and cardiovascular diseases (including in particular heart attacks, heart surgery, and strokes) account for over $\sim 80\%$ of all CII claims, with cancers accounting for the bulk of these.[12]

Major CII markets include Canada, the United Kingdom, Japan, Australia, India, China, and Germany. It is estimated that 20% of British workers were covered by a CII policy in 2009 (Gatzert and Maegebier, 2015). The global CII market has been valued at over $100 billion in 2021 and was projected to grow to over $350 billion by 2031 (Allied Market Research, 2022).

Contract details vary across markets. One dimension of variation is the set of covered illnesses, which has tended to increase over time and is a point of differentiation between competing firms' offerings. Most contracts cover a suite of diseases, though some—such as cancer insurance in the US and Japan (Gatzert and Maegebier, 2015)—only cover one disease or a narrow group of diseases. Another dimension of variation is whether policies include other benefits, such as paying back a fraction of the premium in case of death of the policyholder. Some contracts are standalone, whereas others are riders to another insurance policy such as life insurance; some contracts are purchased via one's employer while others are purchased directly from the insurer.

Underwriting for individual policies is similar to life insurance underwriting. For individual policies, insurers request detailed information, and may ask for medical records, blood or other medical tests, and medical examinations (Canadian Institute of Actuaries, 2012). Industry reports often discuss concerns with adverse selection, and some contract features like waiting periods are meant to curb adverse selection (Gatzert and Maegebier, 2015).

The existing economic literature on the impact of genetic testing on CII is scarce. Our study relates most closely to a series of studies in the actuarial science literature by MacDonald and colleagues (e.g., Macdonald et al., 2005; Macdonald and McIvor, 2009; Macdonald and Tapadar, 2010; Macdonald and Yu, 2011; Adams et al., 2015). They investigate the impact genetic information will have on CII and generally find that adverse selection due to private genetic information is unlikely to be a problem. However, most of these studies are from more than a decade ago and, consistent with the genetic prediction technology at the time, mainly consider single-gene mutations that account for only small shares of the population variation in the studied disorders.

More recently, Maxwell et al. (2021) found that current PGIs predict breast cancer and

[12] In Canada, between 2006 and 2011, 72% and 20% of claims were due to cancers and cardiovascular diseases, respectively (Canadian Institute of Actuaries, 2014). Cancers and cardiovascular diseases together accounted for between 70% and 87% of claims in the US in 2022 (Gen Re Research Center, 2023) and account for over 80% of claims in the UK (Association of British Insurers, 2022).

coronary artery disease above and beyond normal underwriting factors, and discuss the possibility that this may lead to adverse selection. In a study for the Canadian Institute of Actuaries, Howard (2016) considers only some rare genetic variants associated with six conditions and finds that barring insurers from using genetic data would have a sizeable impact on CII claims rates (and thus on premiums) due to adverse selection.

3 Theory

We now introduce our economic and econometric models. We focus on single-disease contracts to facilitate exposition. Section 5.3 and Online Appendix D generalize these models to multiple-disease contracts.

3.1 Economic model

The model. Consider a population of consumers who may incur a binary loss L and can purchase an insurance contract for it. In a single-disease CII contract, the loss corresponds to contracting the disease: $L = D$.

Throughout this section, we consider a population that is treated equally by insurers, such as consumers in the same risk class. We assume that consumers have additional private information, so that there may be selection. This additional private information can include both genetic and non-genetic information.

Each consumer is characterized by the realization of four random variables $(L, G_c, G_f, \mathbf{W})$. The **loss** (or **disease**) L is a random variable equal to zero or one depending on whether a consumer contracts a disease. G_c, the **current polygenic index** (**PGI**) of the disease, is a sufficient statistic of the consumer's DNA for the purpose of estimating the probability of contracting the disease, conditional on the current genetic prediction technology. G_f, the future PGI of L, is a sufficient statistic of the consumer's DNA conditional on the future genetic prediction technology. The **non-genetic covariates W** are a row vector of covariates. The variables $(L, G_c, G_f, \mathbf{W})$ are independently identically distributed across consumers, with joint distribution $\mathbb{P}$. G_c and G_f take real values. $\mathbf{W}$ takes values in Euclidean space.

Private information about risk. Each consumer has private information about her non-genetic factors $\mathbf{W}$ and her PGI(s) G. We will consider four counterfactual scenarios with different sets of consumer private information that vary as a function of the assumed genetic prediction technology: with no genetic prediction technology (Scenario 1); with the current PGI ($G = G_c$; Scenario 2); and with the future PGI ($G = G_f$) with lower-bound

(Scenario 3L) and upper-bound (Scenario 3U) predictive accuracy.

Denote consumers' private information about risk as follows. The **non-genetic private risk function** is defined as the probability $\pi(\mathbf{w})$ of the loss conditional on non-genetic factors $\mathbf{W} = \mathbf{w}$:

$$\pi(\mathbf{w}) := \mathbb{E}\left[L|\mathbf{W} = \mathbf{w}\right].$$

The **private risk function** is defined as the probability $\rho(g, \mathbf{w})$ of the loss conditional on the PGI(s) observed by the consumer, $G = g$, and non-genetic factors $\mathbf{W} = \mathbf{w}$:

$$\rho(g, \mathbf{w}) := \mathbb{E}\left[L|G = g, \mathbf{W} = \mathbf{w}\right].$$

Define the random variable for **non-genetic private risk** as $P := \pi(\mathbf{W})$, and the random variable for **private risk** as $R := \rho(G, \mathbf{W})$. The realization of losses as well as both non-genetic private risk and private risk are independent across consumers.

Measures of adverse selection. The distribution of private risk R across the population is informative about the degree of adverse selection when consumers have genetic information. The left panel of Figure 1 displays the distribution of the risk of contracting prostate cancer by age 65 under the different information scenarios. The blue region represents consumers in the standard risk class. In Scenario 1, where consumers only have non-genetic private risk, the distribution is concentrated around the mean. Therefore, consumers have little private information about their own risk, and consumers who purchase a single-disease CII contract for prostate cancer are not likely to be particularly high-risk, making it easier for an insurance market to perform well. By contrast, the distributions in Scenarios 3L and 3U, where consumers also observe future PGIs, are highly dispersed. Therefore, some consumers know that they have low risk and other consumers know that they have high risk. So consumers who purchase more insurance are likely to be those with particularly high risk. This makes it more likely that an insurance market will perform poorly. In standard adverse selection models, few or no consumers may end up buying insurance, because firms expect that only high-risk consumers would want to purchase and so set a high price, which further deters low-risk consumers, which leads firms to set an even higher price (Akerlof, 1970).

Adverse selection depends on the entire distribution of private risk R. Therefore, for the contracts we study, we will consider the key distributions with histograms like those in Figure 1. Moreover, we will also report an economic measure of selection, the **implicit tax** (Hendren, 2013). The implicit tax $t(r)$ is defined as the extra amount a consumer with private risk r would have to pay if she had to buy insurance at the actuarially fair price for

all consumers with private risk r or greater, versus what her own actuarially fair price is. The formal definition is in Appendix A. Intuitively, the implicit tax tells us how much more the marginal consumer has to pay to buy insurance due to adverse selection, and gives this number as a percentage tax that is easy to interpret.

The implicit tax measure has several advantages. First, Hendren (2013) proposed a no-trade theorem that shows that sufficiently high values of the minimum of the implicit tax cause insurance markets to break down under certain assumptions. Therefore, at least in some settings, the implicit tax is a measure of selection problems.[13] Second, regardless of whether the specific assumptions hold in any given example, the implicit tax is intuitive, and likely to be a relevant statistic even in more general settings.

Finally, we can benchmark the implicit taxes with the values calculated for other markets. Hendren (2013) estimated the implicit tax in market segments that had unraveled (in the sense of not being served by insurers) and in market segments that had not unraveled. He reports the minimum implicit tax up to the 80th percentile of risk, which we denote t_{80}. If t_{80} were to occur at the 70th percentile, say, and were sufficiently low, then insurance in the market segment comprising the top 30% riskiest consumers should be viable. Hendren (2013) considers the *minimum* up to the 80th percentile because if the implicit tax is high at that and at all lower percentiles, then any market segment that comprises the top 20% or more of the riskiest consumers would not be viable, and any segment that comprises less may be too small and thus also unviable. So a high t_{80} is indicative that a market might unravel. For market segments that had not unraveled, Hendren estimated t_{80} to be between 7% and 35%. For market segments that had unraveled, he estimated t_{80} to be between 43% and 83%.

3.2 Econometric model and identification

Our model is fully specified by $\mathbb{P}$, the joint distribution of $(L, G_c, G_f, \mathbf{W})$. While G_c can be observed in the UKB data by the econometrician, G_f cannot. Define the distribution of the observed data as a joint distribution $\mathbb{P}_{\text{data}}$ over $(L, G_c, \mathbf{W})$. We say that the model is identified if there is a unique value of $\mathbb{P}$ consistent with $\mathbb{P}_{\text{data}}$.

We maintain our focus here on the single-disease model, where L is a binary variable equal

[13] A key assumption of this no-trade theorem is that consumers are assumed to have identical risk preferences. However, in equilibrium selection models, the degree of market failure depends both on how risk is distributed and on its joint distribution with risk preferences (Einav et al., 2010a). Section 6.2 further discusses this and shows that the correlation between disease risk and a self-reported measure of risk preferences from the UKB is low. We therefore follow Hendren (2013) in measuring only disease risk distributions. Developing a full structural model for insurance demand and risk would be an interesting topic for future research.

to one if the disease is contracted and G_c and G_f are the disease's current and future PGIs.[14] It may at first appear that the model cannot be identified. Identification implies that we can use the data to estimate the degree of selection in a world in which consumers can observe their future PGIs. It turns out that the model is identified under reasonable assumptions. The key idea is that empirical regularities from quantitative genetics considerably restrict the data, making identification possible. Our assumptions, detailed below, are based on two empirical regularities, which we discussed in Section 2.1. First, PGIs are approximately normally distributed. And second, existing heritability estimates for a disease are informative about the likely predictive power of its future PGI.

We now state the necessary assumptions. The first assumption is a regularity condition on the distribution of covariates.

Assumption 1. *[Covariates' distribution]* $\mathbf{W}$ *is distributed according to a distribution* $\mathbb{P}_{\mathbf{W}}$ *in Euclidean space. The support of* $\mathbf{W}$ *spans the entire space.*

The next two assumptions are conditional normality assumptions on G_c and G_f. They are motivated by the two regularities mentioned above and by the fact that what distinguishes a current PGI from a future PGI for a disease is simply the size of the independent and approximately normally distributed error. Further, because G_c does not have a natural scale, it is standardized to have unit variance and zero mean for each disease in the data.

Assumption 2. *[Gaussian future PGI] Conditional on observables, the future PGI has a Gaussian distribution. The variance is constant and the mean is a linear function of observables:*

$$G_f := \mathbf{W}\boldsymbol{\theta} + V, \tag{1}$$

with $V \sim N(0, \sigma_V^2)$ *and* $V \perp \mathbf{W}$.

Assumption 3. *[Noisy current PGI] The observed current PGI* G_c *has unit variance,* $\mathrm{Var}[G_c] = 1$*, and is a noisy estimate of* G_f*:*

$$G_c := G_f + \epsilon, \tag{2}$$

with $\epsilon \sim N(0, \sigma_\epsilon^2)$ *and* $\epsilon \perp (\mathbf{W}, V)$.

We note that when we assume that a future PGI's accuracy is equal to the disease's twin heritability (as an upper bound), Assumption 3 is likely to be violated. Specifically, in that scenario, it is unlikely to be the case that $\epsilon \perp \mathbf{W}$.Here, we ignore these subtleties, as they

[14]This single-disease econometric model is similar in spirit to that developed by Becker et al. (2021) for continuous traits.

are unlikely to impact our main conclusions. Online Appendix C further discusses this issue and introduces a more complicated model that deals with it, albeit at the cost of additional assumptions.

The fourth assumption is that L follows a probit model, which is the standard functional form for binary diseases in genetic epidemiology. The assumption also implies that the current PGI has no predictive power beyond the future PGI.

Assumption 4. *[Probit disease model] Disease probabilities follow a linear probit model:*

$$\Pr[L = 1 | G_f = g_f, G_c = g_c, \mathbf{W} = \mathbf{w}] = \Phi(g_f \beta_g + \mathbf{w}\boldsymbol{\beta}_w). \tag{3}$$

Under assumptions 1-4, the distribution $\mathbb{P}$ has parameters

$$\Gamma := (\mathbb{P}_{\mathbf{W}}, \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_V^2, \sigma_\epsilon^2).$$

The data can be generated by taking independent draws of $\mathbf{W} \sim \mathbb{P}_{\mathbf{W}}$, $V \sim N(0, \sigma_V^2)$, and $\epsilon \sim N(0, \sigma_\epsilon^2)$. Equation 1 then determines the future PGI G_f; Equation 2 determines the current PGI G_c; and the disease L is realized with probability given by Equation 3.

For the model to be identified, we need one more assumption. We need to assume that we know the predictive power of the disease's future PGI G_f, which we denote R_f^2, from heritability studies. (We can estimate the predictive power of the current PGI, R_c^2, in the data.)

Assumption 5. *[Future PGI predictive power] The pseudo R^2 on the liability scale of a probit regression of L on G_f equals a known constant R_f^2.*

We are now ready to state the identification result.

Theorem 1. *Under assumptions 1-5, the single-disease model is identified.*

Appendix B provides the proof. The intuition for this is as follows. $\mathbb{P}_{\mathbf{W}}$ is identified because $\mathbf{W}$ is observed. The coefficients $\boldsymbol{\theta}$ are the coefficients of a regression of G_c on $\mathbf{W}$. The variances σ_ϵ^2 and σ_V^2 are identified from G_f's assumed future predictive power, the estimated predicted power of G_c in the data, and the variance of $\mathbf{W}\boldsymbol{\theta}$. Finally, the coefficients $\boldsymbol{\beta}$ are related to the coefficients of a probit regression of L on G_c and $\mathbf{W}$. Once the other parameters are known, $\boldsymbol{\beta}$ can be computed from that regression's coefficients.

In Section 5.3 below, we use a generalized version of Theorem 1 that is valid in a multiple-disease-contract setting. Online Appendix D states and proves that generalized version of the theorem.

4 Data and empirical specification

4.1 Data

We analyze data from the UK Biobank (UKB) (Bycroft et al., 2018; Sudlow et al., 2015). The UKB contains genotypic and rich health-related data for over 500,000 individuals from across the United Kingdom who were between 40 and 69 years old at recruitment (between 2006 and 2010). UKB data is linked to the UK's National Health Service (NHS), which maintains detailed records of health events across the lifespan and with which 98% of the UK population is registered (Sudlow et al., 2015). In addition, all UKB participants took part in a baseline assessment, in which they provided rich environmental, family history, health, lifestyle, physical, and sociodemographic data, as well as blood, saliva, and urine samples.

The UKB contains genome-wide array data for ∼800,000 genetic variants for ∼488,000 participants. Our analysis sample comprises the 446,252 UKB participants of European ancestry who have been genotyped and whose genetic data passed a number of quality control filters.[15] Table 2 reports sample descriptive statistics for the seven diseases and the multiple-disease contracts (Panel A) and the diseases' baseline risk factors (Panel B; these are discussed below).

Following standard practices, all genotyping data were imputed (by UKB researchers) using reference genomes. We compute PGIs using these imputed data for the ∼1 million well-characterized and commonly used variants from the HapMap 3 Project (Altshuler et al., 2010; Yengo et al., 2022). For each disease, we use these genetic data and GWAS estimates, together with the software PRS-CS (Ge et al., 2019), to compute the disease's current PGI. The GWAS estimates come from the largest available GWAS of the disease whose analysis sample did not contain the UKB (see Online Appendix A).[16,17] Online Appendix A provides more details on the specific GWASs we used as well as on the quality control filters and how we constructed the PGIs.

The UKB's linked electronic health record (EHR) data cannot be directly analyzed because it is large and unstructured. To code each disease, we rely on the established and widely used "Phecode" mapping system (Denny et al., 2013), which was designed specif-

[15] We focus on the European ancestry participants for the reasons outlined in Section 2.1.

[16] As mentioned above, the GWAS sample must not overlap with the UKB to avoid overfitting (Wray et al., 2013)—i.e., to ensure that $G^* \perp \epsilon$.

[17] As mentioned, Alzheimer's disease and breast cancer are polygenic diseases, but they are additionally affected by variants with sizeable effects in the *APOE* and *BRCA* genes, respectively. In Online Appendix A.6, we show that our PGI for Alzheimer's disease fully captures the *APOE* effects. Our PGI for breast cancer also includes common variants in the *BRCA* genes, but we did not investigate the overlap with rare *BRCA* variants because doing so would require sequencing data.

Table 2: Sample descriptive statistics

Panel A. Descriptive statistics for the diseases

			Age of onset			
Disease	**N**	**% Cases**	**Mean**	**SD**	**5th pct**	**95th pct**
Alzheimer's disease	405,704	0.6%	74.3	74.3	64.4	64.4
Breast cancer (F only)	184,123	6.4%	59.9	59.9	42.2	42.2
Coronary artery disease	410,816	12.0%	62.9	62.9	45.9	45.9
Colorectal cancer	410,646	2.3%	64.9	64.9	48.4	48.4
Prostate cancer (M only)	181,627	7.7%	66.9	66.9	55.4	55.4
Schizophrenia	410,816	0.2%	47.2	47.2	20.5	20.5
Type 2 diabetes	410,816	7.7%	63.7	63.7	47.9	47.9
Multiple-disease (M)	175,148	30.6%	62.8	9.4	46.4	76.5
Multiple-disease (F)	178,128	18.7%	62.2	10.1	44.6	77.0

Panel B. Descriptive statistics for the baseline risk factors (N = 416,261)

Variable	**Mean**	**SD**	**Min**	**Max**
Age (at last observation)	65	9.7	39	85.7
Sex (Male = 1)	0.5	0.5	0	1
Deprivation index (Townsend)	-1.5	2.9	-6.3	10.9
Education (years of schooling)	14	4.9	7	30
BMI	27.4	4.7	12.1	74.7
Drinks per week	8.3	9.3	0	52
Current smoker	0.1	0.3	0	1
Ex smoker	0.4	0.5	0	1
Never smoker	0.5	0.5	0	1
Physical inactivity	0	0.1	0	1
Systolic blood pressure (SBP)	138.3	18.4	72	253.5

Notes: The sample sizes in Panel A reflect the number of participants with no missing values for the disease nor for the baseline or disease-specific risk factors (listed in Table 3). Multiple-disease is a dummy that indicates if one has contracted any of the diseases in the multiple-disease contract.

ically to handle the complexities of defining clinically meaningful case-control status from various complex EHR sources (Bastarache, 2021). The mapping system uses the categories of the International Classification of Diseases (ICD9/10) as reference point.[18]

For each disease, we reviewed the epidemiological literature to identify the main disease-specific non-genetic risk factors. These risk factors are typically used both by medical practitioners and by insurers in underwriting to assess individual disease risk. We then coded variables that capture most of these risk factors in the UKB. Table 3 lists these disease-specific risk factors, together with a set of baseline risk factors which we use for all diseases; the few risk factors we were not able to code in the UKB data are in square brackets. Accounting for conventional epidemiological risk factors in our analysis is critical, because they are typically observed and used by insurers in underwriting. We use these factors, together with a genotyping array dummy and the top ten genetic principal components (PCs) (as is standard in applied genetic research; Price et al., 2006), as the covariates $\mathbf{W}$ in our model.

[18] A key issue addressed by this Phecode mapping system is the high level of heterogeneity among codes in the ICD hierarchy. For example, the general breast cancer ICD10 code C50 (malignant neoplasm of breast) has some ten subcodes to identify what part of the tissue is affected. The Phecode system groups these subtypes together under a single code, 174.1 (Breast cancer [female]). In the latest version, the Phecode

Table 3: Non-genetic risk factors for the studied diseases

Disease	Key risk factors
Baseline risk factors (for all diseases)	Age, sex, Townsend Deprivation index, education (years of schooling), body mass index (BMI), alcohol consumption, smoking (current, ex, or never), physical inactivity, systolic blood pressure, family history* (did each of father, mother, or siblings ever suffer from the disease?)
Alzheimer's disease	Air pollution, [*APOE* risk type], [brain injury], [coronary artery disease], depression, diabetes type 1, [diabetes type 2], hearing loss, hypertension, physical inactivity, social isolation
Breast cancer	Diabetes type 1, [diabetes type 2], ever had breast cancer screening / mammogram, ever had hormonal replacement therapy, hormonal/oral contraceptives, age at menarche, age at menopause, age at first birth, number of live births
Colorectal cancer	Crohn's disease, diabetes type 1, [diabetes type 2], [diet], ever had bowel cancer screening, irritable bowel syndrome (IBS), polyp of colon, ulcerative colitis (and other non-infective gastroenteritis and colitis)
Coronary artery disease	Depression, diabetes type 1, [diabetes type 2], [diet], hypercholesterolemia, hyperglycemia, hypertension, social isolation
Prostate cancer	Ever had prostate specific antigen (PSA) test
Schizophrenia	Anxiety, bipolar disorder, [cannabis use], depression, neuroticism, psychopathology**, [substance use]
Type 2 diabetes	[Diet], hypercholesterolemia, hyperglycemia, hypertension, social isolation

Notes: The non-genetic risk factors were identified in a review of the epidemiological literature and clinical guidelines, and the search findings were verified by two medical doctors. References are reported in Supplementary Table 3. All the risk factors listed in the table were coded as variables in the UKB, with the exception of a few factors (i) that were not clearly defined in the literature (e.g., diet), (ii) for which data was unavailable or insufficient (e.g., brain injury, cannabis use), (iii) that were a disease outcome in the multiple-disease contract (i.e., coronary artery disease and type 2 diabetes), or (iv) that were captured by the PGI (i.e., *APOE*). These risk factors are indicated in square brackets. For more details on the variable coding, see Supplementary Table 4.

* For schizophrenia, family history of "severe depression" was used as a proxy.

** Psychopathology was proxied by neuroticism score, bipolar disorder, and depression.

About half of the risk factors were sourced from the assessment centre data, and the other half from the EHR-data (e.g., comorbidities like hypertension). While the EHR-data is longitudinal and still accruing, most participants had their assessment centre data collected only once, between 2006 and 2010.[19] For the minority of participants with repeated measures, we generally opted for using the most recent non-missing observation (to better match the still accruing EHR-data). For some risk factors, such as BMI, we averaged non-missing repeated observations to reduce measurement error. Also, many of the risk factors vary little over time, or in elderly populations (e.g., highest degree of education, age at first menstruation, or having ever given birth). Our covariate for age was computed from the most recent observation in either the EHR- or assessment centre data. Supplementary Table 4 reports the source of each covariate. Online Appendix B provides additional details and explains how we adjusted age-dependent covariates for age when predicting disease risk by age 65. Overall, the combination of self-reported medical history at the baseline

mapping system condenses some 90,000 ICD-10 codes into about 1,900 Phecodes (Bastarache, 2021).

[19] About 400,000 participants visited an assessment centre only once. Since 2010, a small subset of participants have had up to three repeat assessments: about 20,000 participants were invited in 2012–2013; about 85,000 in 2014–2019; and a final round started in 2019 and has assayed about 8,000 participants so far.

assessment together with longitudinal and still-accruing EHR data let us observe much of the participants' lifetime medical histories.

4.2 Contracts and specification

We consider a loss to be the occurrence of a covered disease. We consider standard contracts that simply pay a lump sum when that loss occurs, thus ignoring the possibility of additional benefits that are sometimes provided by non-standard contracts. Real-world contracts vary across several dimensions (see Section 2.2), so we narrow down our analysis to particular types of contracts, especially in two key contract dimensions: the set of covered diseases and the coverage period. We also specify a population of interest (i.e., a risk class) and the covariates. With this, we have a full specification to take the observable variables from our model (the loss L, the current PGI G_c, and the covariates $\mathbf{W}$) to the data.

Set of covered diseases. CII contracts vary in the number of diseases they cover. Most contracts cover some 20 to 50 illnesses and medical conditions, and the set of diseases has been a point of differentiation between different insurers. Nevertheless, as mentioned, cancers and cardiovascular diseases account for over 80% of all CII claims, with cancers by far the most frequent cause for claims. Some insurers offer enhanced policies with multiple tiers of covered diseases and medical conditions. There are also less comprehensive contracts such as cancer insurance and single-disease CII contracts.

We consider the seven diseases listed in Table 1. These include three of the four most common cancers in Western countries—prostate, breast, and colorectal cancer—which by themselves account for a substantial share of CII claims.[20] As a proxy for cardiovascular disease—the next most common source of claims after cancer—we include coronary artery disease.[21] We include Alzheimer's disease because it is important for long-term care insurance, so the results might be useful for future research on that market. Schizoprenia is included as an example of a highly heritable but rare condition. And we include type 2 diabetes because it is of public health concern and can be included in enhanced CII plans, although it is not covered by standard CII policies.

[20] These three cancers account for a little less than half of all cancers in Western countries. For instance, they accounted for 40% of all cancers in the UK between 2016 and 2018 (Cancer Research UK, 2020). We do not consider the fourth common cancer, lung cancer, because its current PGI has very limited explanatory power over and above the non-genetic covariates; estimating our model with such a PGI would yield imprecise estimates.

[21] Most vasculatory-disease-related CII claims are due to heart attacks (cardiovascular), followed by strokes (cerebrovascular). Heart attacks are a subset of our cardiovascular disease variable and are thus fully captured by it. We do not consider stroke for the same reason that we do not consider lung cancer: the current PGI of stroke has low explanatory power and estimating our model with it would yield imprecise estimates.

Many of the diseases and medical conditions covered by real CII contracts but not by our analysis are rare. Some industry observers believe that insurers include rare diseases because consumers value the more comprehensive policies, perhaps due to poor statistical reasoning, even though the costs are negligible from an actuarial perspective.

We first consider single-disease contracts. While these are not commonly offered as stand-alone contracts, they have some advantages. First, implementation is simpler. Second, these simple contracts are a useful Petri dish to test and illustrate our methodology and they make the interpretation of the results more transparent. In Section 5.3 below, we extend and apply our methodology to multiple-disease CII contracts that bundle the diseases and pay the lump sum upon occurrence of any of the diseases, and then terminate. Since most CII claims are dominated by a few diseases, and since most of these are included in our seven diseases, our multiple-disease contracts approximate real CII contracts.

Coverage period. Typical CII contracts are guaranteed renewable at a predetermined rate up to a maximum age. It is common for the maximum age to be 65. We will restrict attention to this type of contract. The simplest way to model this contract is to assume that all consumers buy contracts at the same age and renew until age 65. Here, we will assume that all consumers buy contracts at age 35, a typical age of first CII purchase that is lower than the usual age of disease onset. Therefore, we will effectively consider one-time 30-year contracts. Ignoring interest rates, this means that insurers' costs are proportional to the probability that the consumer will incur a covered disease by age 65. This is the approximation we will use. In the terminology of our model, we set the loss $L = 1$ if the consumer ever incurs a covered disease by age 65.

We stress that this is an approximation; it departs from reality in two important ways. First, in practice, consumers do let policies lapse. Lapsation is a secondary but important driver of insurance premiums. For example, in life insurance it has been documented that older consumers are more likely to let their policies lapse, even though they have the greatest probability of dying. This is crucial for firms to make positive profits (Gottlieb and Smetters, 2021). Second, the timing of losses also influences insurers' costs, due to time discounting.

Population of interest. We are interested in a population of 35-year-olds who may purchase CII contracts. Our model considers a population of consumers who are treated similarly by insurers. We follow standard underwriting practices to define this group. Insurers often define the **standard risk class** as all consumers who are within 0.75 to 1.25 times the average population-wide risk, based on underwriting using the covariates insurers can observe. It is typical for insurers to charge the same rates to all consumers in the standard risk class. We take the standard risk class to be the population of interest in the analyses

we report in the main text. This fits our model, where we measure selection within a group of consumers who are treated identically by insurers. In addition to the standard risk class, the Online Appendix Figures reports results for a **high risk class** that comprises consumers with non-genetic private risk P between 1.75 and 2.25 times the population average.

PGIs and covariates. For each disease, we construct the current PGI G_c based on the genetic data from the UKB and effect-size estimates from prior GWASs, as described in Section 4.1 and Online Appendix A. Table 3 lists the non-genetic covariates $\mathbf{W}$ for each disease.

Scenarios and information sets. For each contract, we produce results under four Scenarios. In all scenarios, insurers effectively only observe a number of discrete consumer risk classes (including the standard and high risk classes).[22] The scenarios differ with respect to the consumers' private information sets. In Scenario 1, consumers observe only their non-genetic covariates and thus observe no more than a moderate amount of non-genetic private risk P. In the other scenarios, consumers observe both their non-genetic covariates and their current or a future PGI, and thus observe their private risk R. In Scenario 2, consumers observe their current PGI (G_c). In Scenarios 3L and 3U, they observe a future PGI (G_f). In Scenario 3L, the PGI's predictive power is equal to the disease's SNP heritability ($R_f^2 = h_{SNP}^2$), and in Scenario 3U it is equal to the disease's twin heritability ($R_f^2 = h_{twin}^2$).

Scenarios 2 and 3 assume that consumers have widespread access to their genetic risk predictions. As discussed in the Introduction, it is widely believed that this may be the case in the future. Currently, however, there is no widespread consumer access to genetic risk predictions. Therefore, though Scenario 2 is based on the current PGI, it does not reflect the current state of affairs. Section 6.1 below explores scenarios with partial take-up of genetic testing.

4.3 Estimation for single-disease contracts

For all contracts, one minor complication is that, in the UKB, we do not have long-term follow-up data. If we had such follow-up data, we would estimate the model off a population of consumers that we would observe both at age 35 and then again at age 65. The non-genetic covariates would be observed at the former age and disease diagnoses at the latter age. In practice, because of this data limitation, for each disease we use data from individuals of

[22] Though insurers observe all the non-genetic covariates, they only use these to classify consumers across the risk classes, and so they effectively only observe the risk classes. Because the risk classes are somewhat coarse, there is still non-genetic private risk within each class, but not enough to lead to high levels of selection (that's why insurers are content to use these coarse risk classes to set prices).

different ages to estimate the risk each individual develops the disease by age 65.

Section 5.3 and Online Appendix D discuss estimation for the multiple-disease contracts. For each single-disease contracts, estimation is based on the identification Theorem 1, and our method proceeds as follows:

Step 1. Estimating the econometric model. We begin by estimating our econometric model for the contract, using the sample of European ancestry UKB participants who have been genotyped. We pool data from females and males, except for breast cancer and prostate cancer. For Scenarios 1 and 2, estimation involves only a probit regression of the disease on the non-genetic covariates (Scenario 1) and of the disease on both the non-genetic covariates and the current PGI (Scenario 2) (see Appendix B).

For Scenarios 3L and 3U, we first estimate the current PGI's predictive power (R_c^2) in the UKB data, and assume that the future PGI's predictive power (R_f^2) is equal to the disease's SNP heritability and twin heritability for Scenarios 3L and 3U, respectively. Next, we calculate σ_ϵ^2 from Equation B.2 and we estimate the distribution of $\mathbb{P}_W$ as the empirical distribution. Finally, we estimate the remaining parameters $(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_V^2)$ using maximum likelihood estimation, as described in Appendix C.

Step 2. Generating the private risk distributions. For all four scenarios, we generate the distribution of private risk for the disease for all the individuals in our sample. We use Equation 3 along with the parameter estimates obtained from Step 1, the PGI (for Scenarios 2-3), and the age-adjusted covariates (described just below) to estimate each individual's risk of contracting the disease by age 65. For Scenarios 3L and 3U, we do not observe G_f, so for each individual we draw 10 values from its probability distribution conditional on G_c and $\mathbf{W}$ (given by Lemma 1 in the Appendix) and use these to generate the risk distribution.

Individuals in the data range in age from 39 to 86. Our set of covariates $\mathbf{W}$ (see Table 3) includes age as well as a number of age-dependent covariates, such as BMI, blood pressure, hypertension, and whether one ever had menopause (for breast cancer for females) or a prostate specific antigen (PSA) test (for prostate cancer for males). Because age and a number of these covariates are important predictors of risk for all the diseases, we set age to 65 and adjust the most age-dependent covariates to their expected values at age 65 when using Equation 3 to estimate the risk of contracting the disease by age 65. Online Appendix B.3 describes how we adjust these age-dependent covariates for age.

Step 3. Identifying the individuals in the risk classes. To identify the individuals in the standard and high risk classes, we use the risk distribution we generated for Scenario 1. That distribution is based on the non-genetic covariates only, which is what insurers use

to construct the risk classes. Individuals whose risk is between 0.75 times to 1.25 times the average population-wide risk for the disease are assigned to the standard risk class, and those whose risk is between 1.75 to 2.25 times that average risk are assigned to the high risk class.

Step 4. Estimating the implicit tax. For each scenario, we use the scenario's risk distribution from Step 2, limited to the individuals in the standard (or high) risk class. We compute Hendren's implicit tax measure for all percentiles of the distribution. Our main measure of the amount of selection that is likely to ensue is the minimum implicit tax up to the 80th percentile, which we denote t_{80}. We compare that measure with previous estimates for other market segments of the insurance industry. To obtain 95% confidence intervals, we use the bootstrap method, with 199 bootstrap draws.

5 Results

We now discuss our main results. To facilitate exposition, we begin by discussing in detail the results for the single-disease CII contract for prostate cancer and then discuss the main results for all the single-disease CII contracts. We then discuss the results for the multiple-disease contracts.

5.1 Single-disease contracts: case study for prostate cancer

For the single-disease CII contract for prostate cancer, we consider typical male CII buyers aged 35 years old. Each buyer can buy a one-time insurance contract that pays a lump sum if he contracts prostate cancer by age 65.

The risk distributions for prostate cancer under Scenarios 1-3 is illustrated in the left panel of Figure 1 in the Introduction. Consider first Scenario 1, with only non-genetic covariates $\mathbf{W}$. The distribution is bimodal because family history is an important explanatory variable. However, there is still a significant amount of variation due to the other covariates. Overall, there is considerable variation in non-genetic risk. Therefore, there would be a significant amount of selection if insurers had to treat all consumers equally. In practice, insurers underwrite policies based on non-genetic covariates, classify consumers into the aforementioned risk classes, and price these differentially. As mentioned, we focus on consumers in the standard risk class, shown in blue in Figure 1. Since insurers treat these equally, any variation in risk within the risk class effectively is private information. Thus, the blue area in Figure 1, Scenario 1, corresponds to the distribution of non-genetic private risk P for the standard risk class.

We now consider selection under the current genetic prediction technology. Figure 1, Scenario 2, shows in blue the distribution of private risk R (the expectation of a loss conditional on the covariates and the current PGI) for the standard risk class consumers. The current PGI accounts for 11.2% of the liability for prostate cancer, and adds considerable explanatory power over and above the non-genetic covariates: it explains an additional 9.3% of the variance in the liability for prostate cancer, on top of the 23.6% already explained by these covariates. The distribution of private risk is more spread out and is now single-humped because the explanatory power of the coarse family history variable is mostly overshadowed by that of the fine-grained PGI. Thus, even for the consumers in the standard risk class, who look like average risks based on observable covariates, the current PGI introduces significant spread in private risk R. This suggests that, were these consumers to have access to their genetic predictions, there would be a non-negligible amount of selection.

The amount of selection grows considerably with future prediction technology. Figures 1, Scenarios 3L and 3U, plot in blue the distribution of private risk R for the standard risk class based on future PGIs. Scenario 3L uses the lower bound for the future PGI's R^2 based on prostate cancer's SNP heritability, which is 18% (see Table 1). Scenario 3U uses the upper bound based on the twin heritability, which is 58%.

We can use the implicit tax to quantify the amount of selection and compare it to that in previously studied insurance markets. The left panel of Figure 1 plots the implicit taxes in the four scenarios for consumers in the standard risk class. Following Hendren (2013), we focus on the minimum implicit tax up to the 80th percentile, t_{80}. In Scenario 1, with the non-genetic covariates only, t_{80} is equal to 6.8%. This minimum occurs at the 80th percentile.[23] This means that a consumer at that risk percentile would have to pay 6.8% more than his actuarially fair price, if he were charged the actuarial fair price for the segment of consumers with equal or higher risk than his. All consumers at lower risk percentiles face higher implicit taxes. This level of t_{80} has been observed in many well functioning insurance markets (Hendren, 2013), consistent with insurers not using the non-genetic covariates to define narrower risk classes in underwriting.

In Scenario 2, with the current PGI, t_{80} is now 60.6%. This implies that adverse selection would already be significant with today's genetic prediction technology, were all consumers to have routine access to their genetic prediction. This situation is completely different from the current reality. Today, the majority of consumers have never been genotyped and do not have detailed predictions of their risk of contracting prostate cancer and other diseases. Scenario 2 should be interpreted as the amount of selection in a counterfactual with the

[23] In all four scenarios, the implicit tax is a strictly decreasing function of percentile, so the minimum up to the 80th percentile occurs exactly at the 80th percentile.

current prediction technology, but also with widely used genetic prediction. Section 6.1 explores scenarios with only partial take-up of genetic testing.

In Scenario 3L, with a future PGI that explains as much as prostate cancer's SNP heritability, adverse selection increases considerably, with t_{80} now reaching 85.9%. And in Scenario 3U, with a future PGI with predictive power equal to the twin heritability, adverse selection increases dramatically, with t_{80} reaching 452.1%. The implicit taxes in Scenarios 3L and 3U are considerably higher than what Hendren (2013) found even in market segments that had completely unraveled and that insurers refused to cover. This suggests the market for CII for prostate cancer may not be viable in the future if future genetic prediction technology becomes widely available.

The figures also plot the implicit tax at low risk percentiles, where the implicit tax explodes. The implicit tax at the 20th percentile—that for the marginal consumer in the market segment that includes the riskiest 80% of consumers—is very high in all three scenarios with private genetic information. The reason is that, with future genetic prediction, there is a considerable mass of consumers who know they have virtually null risk of ever contracting prostate cancer. To cover most consumers, we need the consumers with negligible risks to pay the price of an average risk. This division by a small number creates the exploding implicit tax.

5.2 Single-disease contracts: main results

We now discuss our results for all seven single-disease CII contracts. Figure 2 reports our estimates of the minimum implicit tax up to the 80th percentile of risk (t_{80}) for each disease and under each scenario for the standard risk class, and Table 4 reports detailed results. Results for the high risk class are shown in Online Appendix Figures C.3-C.10 and are qualitatively similar. Though there are some important nuances, the big picture is broadly consistent with the results for prostate cancer. There are three main findings.

First, the level of selection under current genetic prediction technology and with widespread testing (Scenario 2) would be noticeable. For the standard risk class, the minimum implicit tax t_{80} ranges from 19.7% for coronary artery disease to 84.2% for Alzheimer's. t_{80} for coronary artery disease (19.7%) and colorectal cancer (28.7%) are in the middle of the range found for market segments where there was no unraveling (Hendren, 2013). t_{80} for breast cancer (39.0%) and type 2 diabetes (39.0 and 41.8%) are between the ranges of no unraveling and unravelling. The contracts for the other three diseases—schizophrenia, prostate cancer, and Alzheimer's—have minimum implicit taxes that are all within or above the range observed in market segments that had unraveled. Again, though Scenario 2 assumes today's

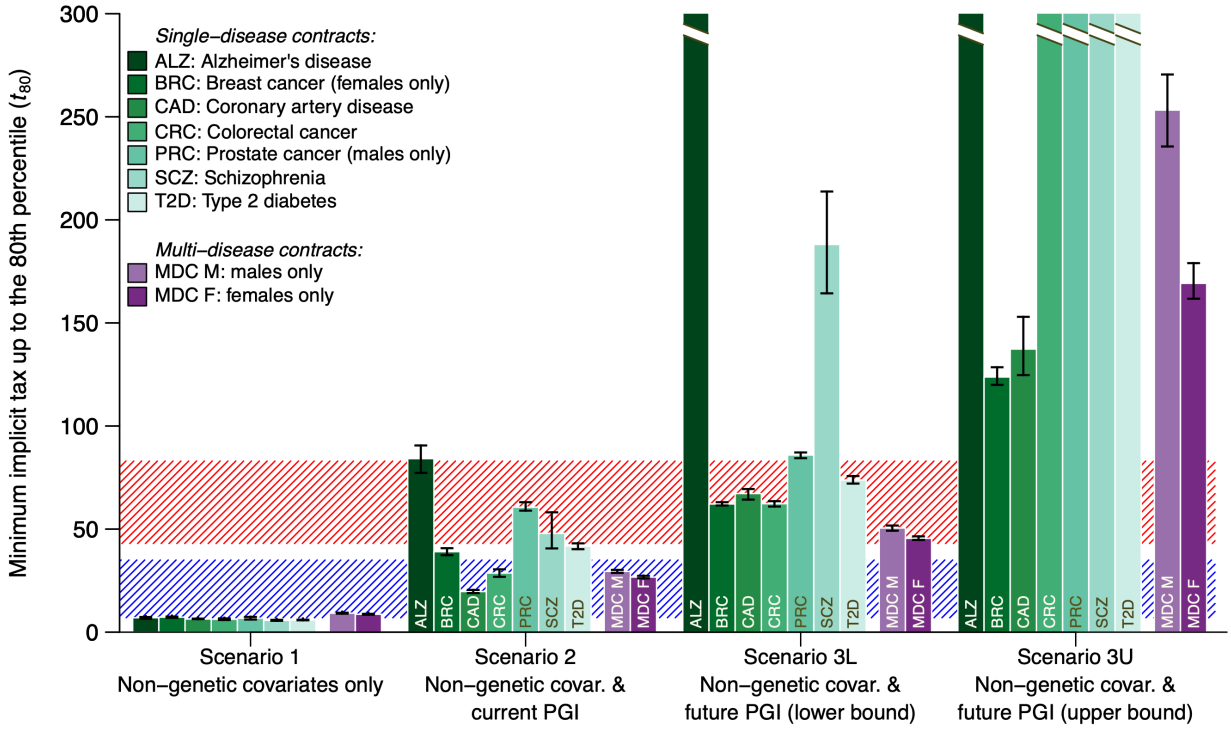

Figure 2: Minimum implicit taxes up to the 80th percentile (t_{80})

Notes: The figure displays t_{80} within the standard risk class for the single-disease and multiple-disease CII contracts, in the four scenarios. The striped blue area corresponds to the range of t_{80} observed by Hendren (2013) in market segments that had not unraveled, and the striped red area corresponds to the range for market segments that had unraveled. The breaks at the top of the tallest bars indicate that these bars' heights exceed the figure's. The results in this figure are also reported in Supplementary Table 7.

genetic prediction technology, it also assumes that all consumers have access to their genetic risk predictions, which does not reflect the situation today. These Scenario 2 results suggest that current prediction technology is already sufficiently accurate to make adverse selection noticeable, and that the main question is when genetic predictions will become widely available.

The second main finding is that, with the expected predictive power of future PGIs, adverse selection will become crippling for many contracts. In Scenario 3L—where PGIs' future R^2's are conservatively assumed to be equal to the diseases' SNP heritabilities—t_{80} ranges from 62.2% for breast and colorectal cancer to >1,000% for Alzheimer's. Some contracts have implicit taxes considerably higher than what Hendren (2013) found even in market segments that had unravelled. The detailed plots in Online Appendix Figures B.3–B.9 show that implicit taxes are high throughout the distribution. The raw histograms of private risk display considerable spread even for consumers who are in the standard risk class. In Scenario 3U—in which future PGIs explain as much as the diseases' twin heritabilities—t_{80}'s are even higher, exceeding 100% for all seven single-disease contracts, and exceeding

1,000% for three. These findings suggest that, with widespread testing and future prediction technology, many single-disease CII contracts will not be viable.

The third main finding is that there is substantial variation between the contracts. The minimum implicit taxes seem to be mainly driven by three factors: the predictive power of the PGI, how much of that predictive power is incremental over that of the non-genetic covariates, and the prevalence of the disease (Table 4). To see this, consider the following examples in Scenario 3L, where the future PGI's predictive power is equal to the SNP h^2. Alzheimer's and schizophrenia have the highest SNP h^2 and have low prevalence. Accordingly, they have the highest implicit taxes. At the opposite end of the spectrum, breast and colorectal cancer have the lowest SNP h^2's and their future PGIs have some of the lowest incremental R^2's in Scenario 3L. Accordingly, these two contracts display the lowest t_{80}'s.

5.3 Multiple-disease contracts

We now consider multiple-disease contracts. We consider separate contracts for males and females that each bundle all the diseases we study, except for breast cancer (for males) and prostate cancer (for females). The multiple-disease contracts pay a lump sum upon occurrence of any of the diseases they cover, after which they terminate. As mentioned, most CII claims are dominated by a few diseases and most of these are included in the diseases we study, so our multiple-disease contracts most closely resemble typical real-world CII contracts. For that reason, we discuss these in detail.

Model and empirical implementation. Our economic model easily generalizes to the multiple-disease setting, except that the **loss** L is now a random variable equal to zero or one depending on whether a consumer contracts any of the diseases listed in the contract. The definitions of non-genetic private risk P and of private risk R still apply. Our econometric model also naturally extends to a multivariate setting, though the details are less straightforward. The reason is that this extension requires specifying and estimating the multivariate distributions of the genetic disturbance terms V, of the current PGI error terms ϵ, and of the probit error across the diseases. Online Appendix D provides these details and states and proves an identification theorem for the a multivariate setting. Estimation follows a similar procedure, also described in Appendix D.

Results. Figure 3 displays the risk distributions and implicit taxes for the multiple-disease contract for males in the standard risk class. The average population risk is 24.3%, but there is a lot of variation in individual risks, even when conditioning only on the non-genetic covariates (Scenario 1). In Scenario 2, when individuals observe both their non-genetic covariates and their current PGIs, the spread of the distribution of private risk R

Table 4: Summary of main results (standard risk class)

	Single-disease contracts							Multiple-disease contracts	
	ALZ	BRC	CAD	CRC	PRC	SCZ	T2D		
Sex	M & F	F only	M & F	M & F	M only	M & F	M & F	M only	F only
Sample size	405,704	184,123	410,816	410,646	181,627	410,816	410,816	175,148	178,128
Cases	2,628	11,696	49,171	9,340	13,926	925	31,756	53,522	33,308
Panel A. Risk of contracting disease (r) by age 65, conditional on non-genetic covariates ($\mathbf{W}$) only									
Full analysis sample									
Mean	0.1%	5.7%	8.2%	1.8%	5.3%	0.2%	5.7%	24.3%	14.6%
SD	0.1 pp	1.7 pp	8.9 pp	1.1 pp	2.1 pp	1.1 pp	8.1 pp	13.2 pp	8.8 pp
Standard risk class (r = 0.75–1.25 of sample mean)									
Mean	0.1%	5.4%	8.0%	1.7%	5.0%	0.2%	5.5%	23.2%	13.8%
SD	0.1 pp	0.7 pp	1.2 pp	0.2 pp	0.8 pp	0.1 pp	0.8 pp	3.4 pp	2.0 pp
Panel B. Explanatory power of non-genetic covariates ($\mathbf{W}$) and PGIs									
Scenarios 1-3									
R^2 of $\mathbf{W}$	36.5%	7.6%	37.3%	12.0%	23.6%	19.8%	40.0%	–	–
Scenarios 2									
R^2 of G_c	6.5%	6.8%	3.0%	2.5%	11.2%	5.8%	8.6%	–	–
R^2 of G_c & $\mathbf{W}$	41.4%	13.7%	38.5%	14.3%	32.9%	23.1%	44.3%	–	–
ΔR^2 of G_c	4.9 pp	6.0 pp	1.2 pp	2.3 pp	9.3 pp	3.4 pp	4.4 pp	–	–
Scenarios 3L									
R^2 of G_f (SNP h^2)	33.1%	14.0%	22.0%	9.0%	18.0%	24.0%	19.6%	–	–
R^2 of G_f & $\mathbf{W}$	63.1%	20.4%	48.1%	20.5%	38.7%	35.9%	50.5%	–	–
ΔR^2 of G_f	26.6 pp	12.7 pp	10.9 pp	8.5 pp	15.2 pp	16.1 pp	10.5 pp	–	–
Scenarios 3U									
R^2 of G_f (twin h^2)	58.0%	31.0%	40.0%	40.2%	57.0%	79.0%	72.0%	–	–
R^2 of G_f & $\mathbf{W}$	85.8%	36.8%	61.8%	55.1%	73.6%	86.2%	91.6%	–	–
ΔR^2 of G_f	49.3 pp	29.2 pp	24.5 pp	43.1 pp	50.0 pp	66.4 pp	51.7 pp	–	–
Panel C. Minimum implicit tax up to the 80th percentile of the disease risk distribution (t_{80})									
No individual underwriting by insurers (same premium for all)									
Scenario 1	75.3%	28.5%	113.1%	42.3%	27.5%	383.0%	134.1%	77.5%	77.2%
Standard risk class (r = 0.75–1.25 of full sample mean)									
Scenario 1	7.0%	7.3%	6.4%	6.3%	6.8%	5.7%	5.9%	9.2%	8.7%
Scenario 2	84.2%	39.0%	19.7%	28.7%	60.6%	47.9%	41.8%	29.5%	26.7%
Scenario 3L	>1,000%	62.2%	67.2%	62.3%	85.9%	188.1%	73.8%	50.5%	45.5%
Scenario 3U	>1,000%	123.8%	137.3%	401.8%	452.1%	>1,000%	>1,000%	253.2%	169.2%

Notes: In Panel A, the reported R^2's are pseudo-R^2's on the liability scale (defined in the text). The R^2 of the non-genetic covariates and of the current PGI (G_c) are estimated in the data; in Scenarios 3L and 3U, the R^2 of the future PGI (G_f) is assumed to be equal to the disease's SNP heritability and twin heritability, respectively, and the joint R^2 of the future PGI and the covariates is obtained from our econometric model. The ΔR^2 of a PGI is its incremental predictive power for the disease, defined as the difference between the R^2 of the PGI and the covariates jointly and the R^2 of the covariates only. References for the heritability estimates are listed in Supplementary Table 1. The four scenarios are defined in the text.

increases. The minimum implicit tax at the 80th percentile, t_{80}, is 29.5% for the standard risk class, implying moderate levels of selection comparable to those observed by Hendren (2013) in markets that had not unraveled. This is consistent with the observation that insurance markets do not yet appear to suffer from elevated levels of selection due to genetic prediction technology (Golinghorst et al., 2022).

In Scenarios 3L and 3U, when individuals observe their future PGIs, the distribution of R becomes very spread out and is almost flat over a large section of its support for the standard risk class. In Scenario 3L, with future PGIs that explain as much as the diseases' SNP heritabilities, consumers at the 5th and 95th percentiles of R face risks of 7.9% and 46.7% of contracting one of the covered diseases; in Scenario 3U, with future PGIs explaining as much as the twin heritabilities, the corresponding figures are 0.95% and 91.7%. Thus, with future genetic prediction technology, the standard risk class contains individuals—all of whom face the same price of insurance—with very low private risk and others with very high private risk, suggesting selection may be an issue. Consistent with this, $t_{80} = 50.5\%$ in Scenario 3L and 253.2% in Scenario 3U, implying elevated levels of selection that have not been observed in well-functioning market segments.

The results for the multiple-disease contract for females in the standard risk class (Online Appendix Figure B.2) and those for males and females in the high risk class (Online Appendix Figures C.1–C.2) are qualitatively very similar. Overall, the broad patterns from the single-disease CII contracts hold for the multiple-disease contracts. In summary, there is noticeable selection with current PGIs (Scenario 2), and serious, potentially crippling selection with future PGIs (Scenarios 3L and 3U). Importantly, multiple-disease contracts display considerably less adverse selection than some of the more extreme single-disease CII. For example, in Scenario 3L, implicit taxes are only marginally in the Hendren's unravelling range for the multiple-disease contracts, but well in the unravelling range for all single-disease contracts. This suggests that bundling may be helpful to reduce selection (though it may not be sufficient to prevent the CII market from unraveling).

6 Robustness and extensions

We now consider several potential limitations to our model and results and examine the robustness of our results to some alternative assumptions.

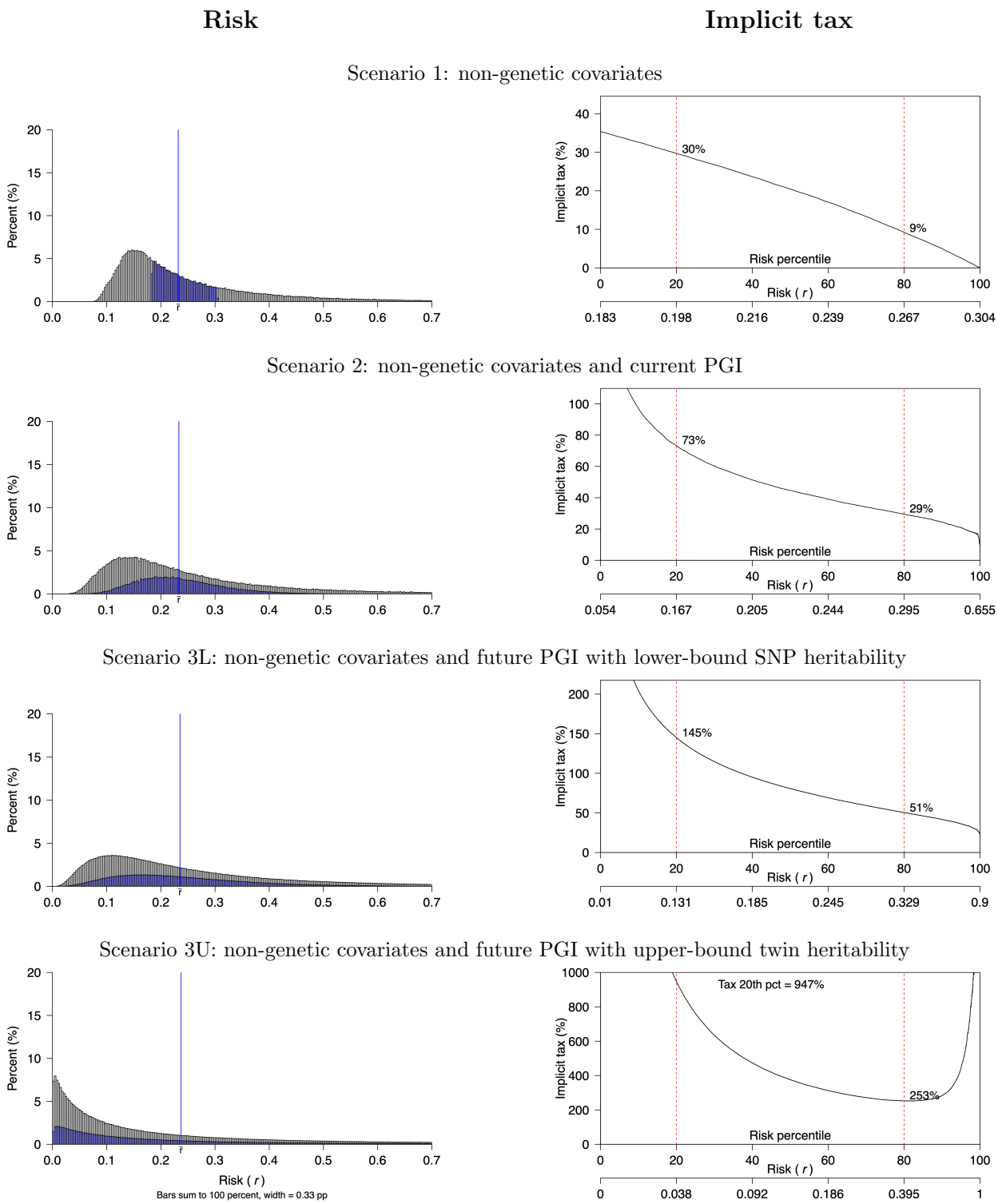

Figure 3: Multiple-disease contract for males: risk and implicit tax

Notes: Left panel: distribution of the risk of contracting any of the six diseases in the male multiple-disease contract by age 65, conditional on various information scenarios and computed in the UKB data ($N = 175,148$). The vertical blue line marks the average risk in the standard risk class ($\bar{r} = 0.233$). The standard risk class individuals are shown in blue; insurers treat these individuals identically, so the blue distribution corresponds to the distribution of private risk for these individuals. Right panel: implicit tax for consumers in the standard risk class as a function of their percentile private risk of contracting one of the diseases, for each scenario.

6.1 Partial take-up of genetic testing

The above results assume that all consumers will perform genetic testing and be aware of their probabilistic predictions. However, there is uncertainty about how common genotyping will become. The modal expert opinion is, arguably, that genotyping will be extremely common.[24] Consistent with this view, as mentioned in the introduction, major consumer genetics companies have started offering disease risk prediction services and large healthcare providers will inform millions of patients of their genetically predicted disease risk.

Nonetheless, to examine the sensitivity of our results to assuming lower take-up of genetic testing, we compute our results under two alternative scenarios: in the Low-take-up Scenario, we assume that only 10% of individuals do genetic tests and take their results into account in their insurance purchase decisions. In the Moderate take-up Scenario, we assume that 50% do genetic tests and take their results into account.

Online Appendix Figures D.1–D.3 show the results. As expected, the amount of selection usually goes down with lower take-up. In particular, in the Low-take-up Scenario—which represents the current-day situation—there would not be much selection regardless of the available prediction technology. In the Moderate take-up Scenario, however, though the minimum implicit taxes t_{80} are lower than under full take-up, they remain problematically high for all the single-disease contracts with future genetic prediction technology (Scenarios 3L and 3U). For the multiple-disease contracts, the minimum implicit taxes t_{80} is just below the problematic region in Scenario 3L and are in the problematic range in Scenario 3U).

Interestingly, incomplete take-up generates kinks and non-monotonicities in the implicit tax function. The main reason is that the risk distributions are not continuous. Consider the Moderate take-up Scenario (see Online Appendix Figure D.1). In Scenarios 2 and 3, the risk histograms are a mixture of the distribution of non-genetic private risk P (for the untested consumers) and the distribution of private risk R (for tested consumers). This induces kinks and non-monotone regions in the implicit tax.

6.2 Advantageous selection

So far, we have measured selection by looking exclusively at the distribution of risk in the population. However, in standard equilibrium models, selection depends both on heterogeneity in risk and in how risk correlates with insurance demand (Akerlof, 1970; Azevedo

[24] The motivation for this standard view is as follows. The cost of genotyping is already of the same order of magnitude as that of a blood test, and has been declining very rapidly. Moreover, genotyping only has to be done once in a patient's lifetime. And, with improved prediction technology, the medical utility of genotyping will be so high that it is anticipated to become integrated in routine clinical care (Wray et al., 2020; Government Office for Science, 2022).

and Gottlieb, 2017). de Meza and Webb (2001) shows that it is possible to have advantageous selection if, for example, low-risk consumers also have higher demand for insurance, possibly because of their higher risk aversion. Fang et al. (2008) find clear evidence of advantageous selection in Medigap insurance. And Finkelstein and McGarry (2006) find that demand for long-term care insurance varies on dimensions unrelated to risk. Therefore, non-risk dimensions of private information can be important, and in principle could affect our conclusions.

Fortunately, the UKB includes a self-reported survey measure of general risk tolerance. Supplementary Table 6 reports various measures of the correlation between general risk tolerance and each of the diseases we study. For all these measures—including for the genetic correlation, which is not attenuated by measurement error[25] —we find extremely low correlations. This suggests that there is no strong negative correlation between disease risk and risk aversion. So it is unlikely that such correlations could have a large impact on our main findings. That said, the UKB risk aversion measure is a limited proxy for insurance demand, so it is possible that other determinants of demand would dampen the amount of selection. But the high levels of selection we find make it unlikely that these other determinants are sufficiently negatively correlated with disease risk to overturn our main qualitative findings.

6.3 Preventive treatment and behavioral changes after testing

One of the main potential benefits of genetic testing is personalized preventive medicine. Genetic tests could be used to identify at-risk patients and offer them preventive cures or inform them of beneficial behavioral changes to mitigate disease risk. Such genetic-testing-informed cures or behavioral changes could in principle decrease the predictive power of future PGIs, since they would lower the actual disease risk of individuals with high PGIs. This would reduce the distributional spread of private risk and thus lower the minimum implicit tax t_{80} and reduce selection. In practice, this effect is unlikely to substantively affect our main findings. We find very high levels of selection for most contracts. So, given plausible assumptions about the take-up of preventive actions and the effectiveness of these actions, this is unlikely to reduce selection so much as to overturn our main findings. This is especially true of cancers, because known preventive actions have limited efficacy.

[25] We estimate genetic correlations with LD Score regressions (Bulik-Sullivan et al., 2015), which yields estimates that are not attenuated by measurement error. Under some assumptions, these estimates are equal to the correlation between the variables' true additive genetic factors.

6.4 Firms with an informational advantage

There are a few theoretical papers about insurance markets that assume insurers have private information.[26] Genetic testing could make such theoretical assumptions plausible in some institutional settings. For example, consider a future world where insurers are allowed to underwrite based on genetics, tests are widely available, but consumers do not have the sophistication to interpret the tests. This would correspond to a model where insurers receive private information but consumers do not.

The simplest and most robust prediction of this type of model is in the competitive case. If insurers just compete their profits to zero, they will offer actuarially fair insurance to each consumer. Thus, even if consumers can't interpret their genetic tests, they will be able to assess their risk by getting insurance quotes. In such a future world, there is effectively no private information and equilibrium is the same regardless of whether consumers can or cannot interpret their genetic tests. Equilibrium is efficient, in the sense that each consumer is charged her true cost for insurance. But equilibrium can be seen as very unfair, as a consumer who is unlucky in the genetic lottery will face high prices.

7 Policy implications

Many developed countries already ban insurers from using genetic information and others are moving in that direction (Nabholz and Rechfeld, 2017; Middleton, 2024). Our findings suggest that, with future genetic prediction technology and if genetic information becomes widely available, these bans will lead to a high degree of adverse selection in CII, with the risk of complete unravelling of the market for most of the contracts we study.

Optimal policy responses. While genetic predictions are new, selection is a standard issue in insurance markets. There is a large amount of research and practical experience on how to regulate selection markets to address market failures due to asymmetric information (Einav et al., 2010b, 2021, 2023). The problem of regulating genetic information in insurance is similar. Therefore, we can draw on extensive research and practical experience from other selection markets. Online Appendix E gives a brief overview of the literature and of standard policy instruments.

In a nutshell, regulation of selection markets typically seeks to balance two goals: efficiency and redistribution. Efficiency aims to maximize total economic surplus. Redistribution aims to help particular groups, in particular high risk groups who face a higher price of insurance (but also consumers with lower wealth or worse health status). The standard regu-

[26] See Brunnermeier et al. (2021) and the references therein.

latory playbook can be divided in three approaches: laissez-faire, government provision, and managed competition. Here, laissez-faire would involve allowing insurers to observe genetic information and to use it in underwriting, without governmental intervention. This would result in the highest risk consumers facing very high premiums and may be deemed unfair. Furthermore, there is widespread concern that this could lead many to forego genetic testing (and the associated medical benefits) or to share their genetic data for research, for fear that their test results could lead to increased premiums. As for government provision may not be attractive for a product such as CII that is not seen as essential for many consumers.

Empirical work similar to existing research on other selection markets will be necessary to design optimal policies for the CII market. Nonetheless, a natural policy response that falls under the rubric of managed competition could involve a mix of community rating, subsidies, and risk adjustment. Community rating are rules that limit price discrimination. The current ban on genetic underwriting is a community rating regulation. Community rating improves redistribution across consumers, but generates adverse selection. Some level of community rating may be desirable to help high-risk consumers, but a full ban on genetic price discrimination may be too much due to the large high levels of adverse selection it would cause. Likewise, subsidies might be desirable to increase efficiency and help vulnerable populations, at the cost of spending tax revenue. Finally, risk adjustment are cross subsidies to insurers based on the populations they enroll may also help.

In addition to likely leading to selection issues (as shown in this study), a full ban on genetic underwriting may hinder the development of welfare-improving initiatives by insurers. For instance, if allowed to use genetic data, insurers could give discounts to incentivize high genetic risk consumers to undertake preventive treatments or behavioral changes to reduce their risk.

Political economy issues. Improving on the status quo policy of banning insurers from using genetic information is politically complex. For example, the existing bans on insurers' use of genetic testing were enacted with the aim of protecting consumers. However, if a market such as CII unravels, all consumers will be worse off, because they will not get any insurance at all. A regulation that allows insurers to underwrite based on genetic information would benefit all consumers—including the high risk ones—because they would then at least be able to get some insurance. However, high-risk consumers may be upset about the high premiums they would have to pay. Thus, even consumers who would benefit from the regulation would probably oppose it after the fact, unless it were accompanied by other policy tools like the ones discussed above to improve redistribution.

Policies specific to genetic testing. Some existing policy proposals are specifically

aimed at regulating genetic information in insurance. The most well-known is Tabarrok's (1994) idea of genetic insurance. In a world where insurers are allowed to use genetic information, Tabarrok proposes an insurance contract that would be mandatory for anyone taking a genetic test for the first time. Tabarrok considered genetic tests that gave clear positive or negative results for a handful of rare diseases caused by one or a few genes. Upon a positive genetic test result, a consumer would get paid a lump sum equal to the expected cost of the disease (or, equivalently, to the new expected health insurance premium). That way, high risk consumers can still purchase insurance even though insurers observe their genetics. Tabarrok shows that, under certain assumptions, this policy is second-best optimal.

Tabarrok's idea can be extended naturally to our setting with CII contracts for common polygenic diseases with continuously distributed PGIs. The mandatory genetic insurance for all first-time genetic testers would pay one's expected premium conditional on one's PGIs for a pre-specified CII contract. However, unlike for the rare diseases Tabarrok considered, the premiums for such mandatory genetic insurance for CII would be large. Consider genetic insurance for our multi-disease CII contract for males with a payment of $100,000 upon the onset of a covered disease. Before taking a genetic test, each consumer would have to pay a premium of $24,300, based on the average probability of 24.3% of contracting at least one of the covered diseases. It may be politically fraught to ask consumers to pay such large sums before taking a genetic test for the first time. Moreover, many consumers never purchase CII contracts, and it would be difficult to agree on a set of diseases to be covered and on the covered amounts. Further, some consumers may secretly take genetic tests and never report doing this.

Another possibility is to consider industry-led solutions. For example, one idea is to allow consumers to disclose that they have not taken a genetic test at the time of purchasing a CII policy. This could in principle allow at least some consumers to purchase CII with little genetic adverse selection. For this idea to be feasible, however, insurers would have to be able to detect and then void the contracts of consumers who fail to disclose they got tested.

8 Conclusion

We developed a methodology to measure adverse selection in insurance markets due to genetic prediction of common diseases. We push the research frontiers along several dimensions. First, we consider genetic prediction based on PGIs, which depend on many common genetic variants, as opposed to diagnostic genetic tests for rare mutations. Second, we develop a new method that uses heritability estimates as reasonable bounds for the future predictive power of genetic prediction, and considers the incremental predictive power of genetic prediction

over and above observable risk factors. This allows us to estimate how selection will change as prediction technology improves. And third, using the UKB's comprehensive, large-scale micro-data with rich information on non-genetic risk factors, genotypes, and diseases, we apply our method to the CII market. We find that, if genetic testing becomes widespread, there would be noticeable selection with the current prediction technology. The amount of selection would be crippling with the expected future prediction technology.

Genetic prediction is a powerful new technology that has the potential to bring about important benefits, such as personalized medicine and preventive treatments. Understanding the unintended consequences of this technology can help society mitigate its negative effects while still harvesting its benefits. This study aims to help understand some of these possible unintended consequences in the realm of insurance markets. Future work should seek to address the above limitations and to apply our methodology to larger markets like life, health, and long-term care insurance.

References

23andMe. Information about Genetic Health Risk Reports—23andMe. https://www.23andme.com/test-info/genetic-health/, 2023.

Abdellaoui, A., Yengo, L., Verweij, K. J., and Visscher, P. M. 15 years of GWAS discovery: Realizing the promise. *The American Journal of Human Genetics*, 2023.

Adams, C., Donnelly, C., and Macdonald, A. The impact of known breast cancer polygenes on critical illness insurance. *Scandinavian Actuarial Journal*, 2015(2):141–171, 2015.

Akerlof, G. The market for "lemons": quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84(3):488–500, 1970.

Allied Market Research. Critical illness insurance market, 2022.

Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467:52–58, 2010.

Association of British Insurers. ABI guide to minimum standards for criticall illness cover. *Mimeo*, 2022.

Azevedo, E. M. and Gottlieb, D. Perfect competition in markets with adverse selection. *Econometrica*, 85(1):67–105, 2017.

Bastarache, L. Using phecodes for research with the electronic health record: from PheWAS to PheRS. *Annual Review of Biomedical Data Science*, pages 1–19, 2021.

Becker, J., Burik, C. A., Goldman, G., Wang, N., Jayashankar, H., et al. Resource profile and user guide of the polygenic index repository. *Nature human behaviour*, 5(12):1744–1758, 2021.

Bélisle-Pipon, J.-C., Vayena, E., Green, R. C., and Cohen, I. G. Genetic testing, insurance discrimination and medical research: what the United States can learn from peer countries. *Nature Medicine*, 25(8):1198–1204, 2019.

Brackenridge, R. D. C., Croxson, R. S., and Mackenzie, R., editors. *Medical selection of life risks*. Palgrave Macmillan, New York, fifth edition, 2006.

Brunnermeier, M. K., Lamba, R., and Segura-Rodriguez, C. Inverse selection. *SSRN paper no. 3584331*, 2021.

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., et al. An atlas of genetic correlations across human diseases and traits. *Nature genetics*, 47(11):1236–1241, 2015.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

Canadian Institute of Actuaries. Critically Canadian: Canadian critical illness standalone base incidence tables. *Mimeo*, 2012.

Canadian Institute of Actuaries. Canadian individual critical illness insurance morbidity experience study. *Mimeo*, 2014.

Cancer Research UK. Cancer incidence for common cancers. https://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence/common-cancers-compared#

heading-Zero, 2020. Accessed: 2024-06-06.

Chabris, C. F., Lee, J. J., Cesarini, D., Benjamin, D. J., and Laibson, D. I. The fourth law of behavior genetics. *Current Directions in Psychological Science*, 24(4):304–312, 2015.

Daetwyler, H. D., Villanueva, B., and Woolliams, J. A. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PloS One*, 3(10):e3395, 2008.

de Meza, D. and Webb, D. C. Advantageous selection in insurance markets. *RAND Journal of Economics*, 32(2):249–262, 2001.

Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology*, pages 1102–1111, 2013.

Dixon, P., Horton, R. H., Newman, W. G., McDermott, J. H., and Lucassen, A. Genomics and insurance in the United Kingdom: increasing complexity and emerging challenges. *Health Economics, Policy and Law*, pages 1–13, 2024.

Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genetics*, 9: e1003348, 2013.

Einav, L., Finkelstein, A., and Cullen, M. R. Estimating welfare in insurance markets using variation in prices. *The Quarterly Journal of Economics*, 125(3):877–921, 2010a.

Einav, L., Finkelstein, A., and Levin, J. Beyond testing: empirical models of insurance markets. *Annu. Rev. Econ.*, 2(1):311–336, 2010b.

Einav, L., Finkelstein, A., and Mahoney, N. The IO of selection markets. In *Handbook of Industrial Organization*, volume 5, pages 389–426. Elsevier, 2021.

Einav, L., Finkelstein, A., and Fisman, R. *Risky business: why insurance markets fail and what to do about it*. Yale University Press, 2023.

Falconer, D. S. and Mackay, T. F. *Introduction to quantitative genetics*. Prentice Hall, Essex, 4th edition, 1996.

Fang, H., Keane, M. P., and Silverman, D. Sources of advantageous selection: evidence from the medigap insurance market. *Journal of Political Economy*, 116(2):303–350, 2008.

Finkelstein, A. and McGarry, K. Multiple dimensions of private information: evidence from the long-term care insurance market. *American Economic Review*, 96(4):938–958, 2006.

Gatzert, N. and Maegebier, A. Critical illness insurances: challenges and opportunities for insurers. *Risk Management and Insurance Review*, 18(2):255–272, 2015.

Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., and Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications*, 10:1776, 2019.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. *Bayesian data analysis*. Chapman and Hall/CRC, 3rd edition, 2013.

Gen Re Research Center. U.S. critical illness and accident insurance market survey. *Mimeo*, 2023.

Golinghorst, D., De Paor, A., Joly, Y., Macdonald, A. S., Otlowski, M., et al. Anti-selection & genetic testing in insurance: an interdisciplinary perspective. *Journal of Law, Medicine*

& Ethics, 50(1):139–154, 2022.

Gottlieb, D. and Smetters, K. Lapse-based insurance. *American Economic Review*, 111(8): 2377–2416, 2021.

Government Office for Science. Genomic beyond health: what could genomics mean for wider government? Technical report, Government Office for Science, 2022.

Hendren, N. Private information and insurance rejections. *Econometrica*, 81(5):1713–1762, 2013.

Hendren, N. Knowledge of future job loss and implications for unemployment insurance. *American Economic Review*, 107(7):1778–1823, 2017.

Hill, W. G., Goddard, M. E., and Visscher, P. M. Data and theory point to mainly additive genetic variance for complex traits. *PLOS Genetics*, 4(2):e1000008, 2008.

Holland, D., Frei, O., Desikan, R., Fan, C.-C., Shadrin, A. A., et al. Beyond SNP heritability: polygenicity and discoverability of phenotypes estimated with a univariate Gaussian mixture model. *PLOS Genetics*, 16(5):e1008612, 2020.

Howard, R. C. Genetic testing model for CI: if underwriters of individual critical illness insurance had no access to known results of genetic tests. *Canadian Institute of Actuaries Report to CIA Research Committee*, 2016.

Joly, Y., Burton, H., Knoppers, B. M., Feze, I. N., Dent, T., et al. Life insurance: genomic stratification and risk classification. *European Journal of Human Genetics*, 22:575–579, 2014.

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(9):1219–1224, 2018.

Klarin, D. and Natarajan, P. Clinical utility of polygenic risk scores for coronary artery disease. *Nature Reviews Cardiology*, 19(5):291–301, 2022.

Lee, S. H., Goddard, M. E., Wray, N. R., and Visscher, P. M. A better coefficient of determination for genetic profile analysis. *Genetic Epidemiology*, 36(3):214–224, 2012.

Loos, R. J. 15 years of genome-wide association studies and no signs of slowing down. *Nature Communications*, 11(1):1–3, 2020.

Macdonald, A. The actuarial relevance of genetic information in the life and health insurance context. Technical report, Office of the Privacy Commissioner of Canada, 2011.

Macdonald, A. and Tapadar, P. Multifactorial genetic disorders and adverse selection: epidemiology meets economics. *Journal of Risk and Insurance*, 77(1):155–182, 2010.

Macdonald, A. and Yu, F. The impact of genetic information on the insurance industry: conclusions from the "bottom-up" modelling programme. *ASTIN Bulletin*, 41(2):343–376, 2011.

Macdonald, A. S. and McIvor, K. R. Modelling adverse selection in the presence of a common genetic disorder: the breast cancer polygene. *ASTIN Bulletin*, 39(2):373–402, 2009.

Macdonald, A. S., Waters, H. R., and Wekwete, C. T. A model for coronary heart disease and stroke with applications to critical illness insurance underwriting I: the model. *North*

American Actuarial Journal, 9(1):13–40, 2005.

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., et al. Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics*, 100(4):635–649, 2017.

Maxwell, J. M., Russell, R. A., Wu, H. M., Sharapova, N., Banthorpe, P., et al. Multifactorial disorders and polygenic risk scores: predicting common diseases and the possibility of adverse selection in life and protection insurance. *Annals of Actuarial Science*, 15(3): 488–503, 2021.

McKelvey, R. D. and Zavoina, W. A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4(1):103–120, 1975.

Meyer, M. N., Papageorge, N. W., Parens, E., Regenberg, A., Sugarman, J., et al. Potential corporate uses of polygenic indexes: starting a conversation about the associated ethics and policy issues. *The American Journal of Human Genetics*, 111:833–840, 2024.

Middleton, K. Australia to ban life insurance companies from discriminating based on genetic testing results. *The Guardian*, 2024.

Mills, M. C. and Rahal, C. The GWAS Diversity Monitor tracks diversity by disease in real time. *Nature Genetics*, 52(3):242–243, 2020.

MyHeritage. Understanding your polygenic risk reports. https://education.myheritage.com/article/understanding-your-polygenic-risk-reports/, 2021.

Nabholz, C. and Rechfeld, F. Seeing the future? How genetic testing will impact life insurance. Technical report, Swiss Re Institute, 2017.

O'Connor, L. J., Schoech, A. P., Hormozdiari, F., Gazal, S., Patterson, N., et al. Extreme polygenicity of complex traits is explained by negative selection. *The American Journal of Human Genetics*, 105:456–476, 2019.

Okbay, A., Wu, Y., Wang, N., Jayashankar, H., Bennett, M., et al. Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nature Genetics*, 54(4):437–449, 2022.

Oster, E., Shoulson, I., Quaid, K., and Dorsey, E. R. Genetic adverse selection: evidence from long-term care insurance and huntington disease. *Journal of Public Economics*, 94 (11-12):1041–1050, 2010.

Our Future Health. Our Future Health supports Genome UK implementation plan for England, 2022.

Polderman, T. J. C., Benyamin, B., de Leeuw, C., Sullivan, P. F., van Bochoven, et al. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, 47(7):702–709, 2015.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

Privé, F., Aschard, H., Carmi, S., Folkersen, L., Hoggart, C., et al. Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups

from the same cohort. *The American Journal of Human Genetics*, 109(1):12–23, 2022.

Rechfeld, F., Rischatsch, M., Nemoto, T., and Nabholz, C. Can life insurance pass the genetic test? Technical report, Swiss Re Institute, Zurich, 2019.

Regalado, A. More than 26 million people have taken an at-home ancestry test, 2019.

Saya, S., Boyd, L., Chondros, P., McNamara, M., King, M., et al. The SCRIPT trial: study protocol for a randomised controlled trial of a polygenic risk score to tailor colorectal cancer screening in primary care. *Trials*, 23(1):810, 2022.

Simcoe, M., Valdes, A., Liu, F., Furlotte, N. A., Evans, D. M., et al. Genome-wide association study in almost 195,000 individuals identifies 50 previously unidentified genetic loci for eye color. *Science Advances*, 7(11):eabd1239, 2021.

So, H.-C. and Sham, P. C. A unifying framework for evaluating the predictive power of genetic variants based on the level of heritability explained. *PLoS Genetics*, 6(12):e1001230, 2010.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, 12(3):e1001779, 2015.

Tabarrok, A. Genetic testing: an economic and contractarian analysis. *Journal of Health Economics*, 13(1):75–91, 1994.

Taylor Jr, D. H., Cook-Deegan, R. M., Hiraki, S., Roberts, J. S., Blazer, D. G., et al. Genetic testing for Alzheimer's and long-term care insurance. *Health Affairs*, 29(1):102–108, 2010.

Torkamani, A., Wineinger, N. E., and Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19:581—590, 2018.

Visscher, P. M., Yengo, L., Cox, N. J., and Wray, N. R. Discovery and implications of polygenicity of common diseases. *Science*, 373:1468–1473, 2021.

Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E., et al. Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics*, 14(7):507–515, 2013.

Wray, N. R., Lin, T., Austin, J., McGrath, J. J., Hickie, I. B., et al. From basic science to clinical application of polygenic risk scores: a primer. *JAMA Psychiatry*, 78:101–109, 2020.

Yanes, T., Tiller, J., Haining, C. M., Wallingford, C., Otlowski, M., et al. Future implications of polygenic risk scores for life insurance underwriting. *npj Genomic Medicine*, 9:25, 2024.

Yengo, L., Vedantam, S., Marouli, E., Sidorenko, J., Bartell, E., et al. A saturated map of common genetic variants associated with human height. *Nature*, 610:704–712, 2022.

Zeggini, E., Gloyn, A. L., Barton, A. C., and Wain, L. V. Translational genomics and precision medicine: moving from the lab to the clinic. *Science*, 365(6460):1409–1413, 2019.

Zhou, W., Kanai, M., Wu, K.-H. H., Rasheed, H., Tsuo, K., et al. Global biobank meta-analysis initiative: powering genetic discovery across human disease. *Cell Genomics*, 2 (10), 2022.

APPENDIX

A Definition of the implicit tax

Consider an insurance contract that pays \$1 if a loss happens, and otherwise charges a premium. For the contract to break even with a consumer with private risk r, the premium has to be

$$\frac{r}{1-r}.$$

Consider now the case where this policy is purchased by all consumers with private risk r or greater. In that case, for the policy to break even the premium would have to be

$$\frac{\mathbb{E}[R|R \geq r]}{1-\mathbb{E}[R|R \geq r]}.$$

The consumer with private risk r would have to pay $T(r)$ times the actuarially fair price, where

$$T(r) := \frac{\mathbb{E}[R|R \geq r]}{1-\mathbb{E}[R|R \geq r]} \Big/ \frac{r}{1-r}.$$

Hendren (2013) calls $T(r)$ the "pooled price ratio". To make exposition more intuitive, we define the **implicit tax** $t(r)$ as

$$1 + t(r) := T(r). \tag{A.1}$$

So $t(r)$ is the implicit tax that a consumer with risk r would have to pay due to selection if she were charged an actuarially fair price for all higher-risk consumers.

B Proof of Theorem 1

Here we prove that the single-disease econometric model is identified under Assumptions (1)-(5) (Theorem 1). The proof uses the following Lemma.

Lemma 1. *Conditional on $G_c = g_c$ and $\mathbf{W} = \mathbf{w}$, G_f is normally distributed with mean*

$$ag_c + b\mathbf{w}\boldsymbol{\theta}$$

and variance c^2, where

$$a = \frac{\frac{1}{\sigma_\epsilon^2}}{\frac{1}{\sigma_\epsilon^2} + \frac{1}{\sigma_V^2}},$$

$$b = \frac{\frac{1}{\sigma_V^2}}{\frac{1}{\sigma_\epsilon^2} + \frac{1}{\sigma_V^2}},$$

$$\frac{1}{c^2} = \frac{1}{\sigma_\epsilon^2} + \frac{1}{\sigma_V^2}.$$

Proof. By Assumption 2, conditional on $\mathbf{W} = \mathbf{w}$, $G_f = \mathbf{w}\boldsymbol{\theta} + V$, with $V \sim N(0, \sigma_V^2)$. And by Assumption 3, $G_c = G_f + \epsilon$, with $\epsilon \sim N(0, \sigma_\epsilon^2)$. Lemma 1 then follows from the formula in Gelman et al. (2013, p. 40, Equation 2.10). This is a standard formula from Bayesian statistics. □

Proof of Theorem 1. We will show that the parameters $\Gamma = (\mathbb{P}_{\mathbf{W}}, \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_V^2, \sigma_\epsilon^2)$ are functions of $\mathbb{P}_{\text{data}}$. $\mathbb{P}_{\mathbf{W}}$ is trivial because $\mathbf{W}$ is observed. We now turn to the other parameters.

Part 1: $\boldsymbol{\theta}$

Assumption 4 implies that

$$\mathbb{E}[G_c|\mathbf{W} = \mathbf{w}] = \mathbf{w}\boldsymbol{\theta}. \tag{B.1}$$

Assumption 1 implies that there is a set of $\mathbf{w}$'s in the support of $\mathbb{P}_{\mathbf{W}}$ that span the entire space. So $\boldsymbol{\theta}$ is the vector of coefficients of a regression of G_c on $\mathbf{W}$.

Part 2: σ_ϵ^2

By Assumption 4, there exists a latent variable $\mathcal{L}$ for the disease such that

$$\mathcal{L} := G\beta_g + \boldsymbol{W}\boldsymbol{\beta}_w + \eta,$$

where $\eta \sim N(0, 1)$, $\eta \perp (\boldsymbol{W}, V, \epsilon)$, and $L = \{\mathcal{L} > 0\}$, and where the operator $\{..\}$ is equal to 1 if its argument is true and to 0 otherwise.

Let the pseudo-R^2 of a probit regression of L on the current PGI be R_c^2. R_c^2 is a function of the data. R_c^2 equals

$$R_c^2 = \frac{\text{Cov}[\mathcal{L}, G_c]^2}{\text{Var}[G_c]\text{Var}[\mathcal{L}]}.$$

The pseudo-R^2 of a probit regression of L on the future PGI is

$$R_f^2 = \frac{\text{Cov}[\mathcal{L}, G_f]^2}{\text{Var}[G_f]\text{Var}[\mathcal{L}]}.$$

Therefore, using the fact that $\text{Cov}[\mathcal{L}, G_f] = \text{Cov}[\mathcal{L}, G_c]$ and $\text{Var}[G_c] = 1$,

$$\frac{R_c^2}{R_f^2} = \frac{\text{Var}[G_f]}{\text{Var}[G_c]} = \text{Var}[G_f].$$

Assumption 3 implies that

$$\text{Var}[G_c] = \text{Var}[G_f] + \sigma_\epsilon^2.$$

Plugging in the formula above we have

$$\sigma_\epsilon^2 = 1 - \frac{R_c^2}{R_f^2}. \tag{B.2}$$

Part 3: σ_V^2

Assumptions 3 and 2 yield

$$G_c = \mathbf{W}\boldsymbol{\theta} + V + \epsilon. \tag{B.3}$$

Taking the variance we find

$$1 = \mathrm{Var}[\mathbf{W}\boldsymbol{\theta}] + \sigma_V^2 + \sigma_\epsilon^2,$$

so that

$$\sigma_V^2 = 1 - \mathrm{Var}[\mathbf{W}\boldsymbol{\theta}] - \sigma_\epsilon^2. \tag{B.4}$$

Part 4: β_g

As mentioned in Part 2, by Assumption 4, we have

$$L = \{G_f\beta_g + \mathbf{W}\boldsymbol{\beta}_w + \eta > 0\}.$$

Consider the distribution of L conditional on $G_c = g_c$ and $\mathbf{W} = \mathbf{w}$. Lemma 1 implies that

$$G_f = ag_c + b\mathbf{w}\boldsymbol{\theta} + c\nu,$$

where ν is a standard normal random variable independent of η. Therefore, conditional on $G_c = g_c$ and $\mathbf{W} = \mathbf{w}$,

$$\begin{aligned} L &= \{(ag_c + b\mathbf{w}\boldsymbol{\theta} + c\nu)\beta_g + \mathbf{w}\boldsymbol{\beta}_w + \eta > 0\} \\ &= \{g_c a\beta_g + \mathbf{w}(b\boldsymbol{\theta}\beta_g + \boldsymbol{\beta}_w) + \eta + c\nu\beta_g > 0\} \\ &= \left\{ g_c \frac{a\beta_g}{\sqrt{1+c^2\beta_g^2}} + \mathbf{w}\frac{b\boldsymbol{\theta}\beta_g + \boldsymbol{\beta}_w}{\sqrt{1+c^2\beta_g^2}} + \frac{\eta + c\nu\beta_g}{\sqrt{1+c^2\beta_g^2}} > 0 \right\}. \end{aligned}$$

The last term is distributed as $N(0,1)$. Therefore

$$\Pr(L = 1 | G_c = g_c, \mathbf{W} = \mathbf{w}) = \Phi\left[g_c \frac{a\beta_g}{\sqrt{1+c^2\beta_g^2}} + \mathbf{w}\frac{b\boldsymbol{\theta}\beta_g + \boldsymbol{\beta}_w}{\sqrt{1+c^2\beta_g^2}} \right]. \tag{B.5}$$

Let $\boldsymbol{\gamma}$ be the coefficient vector of a probit regression of L on G_c and $\mathbf{W}$. Then

$$\frac{a\beta_g}{\sqrt{1+c^2\beta_g^2}} = \gamma_g.$$

This equation implies that β_g has the same sign as γ_g. Taking the square of both sides and solving, we find

$$\beta_g = \frac{\gamma_g}{\sqrt{a^2 - \gamma_g^2 c^2}}. \tag{B.6}$$

This result implies that $|\gamma_g| < a/c$.

Part 5: $\boldsymbol{\beta}_w$

By Equation B.5 and using the fact that the support of $\mathbf{W}$ includes a set that spans the entire space (Assumption 1),

$$\frac{b\boldsymbol{\theta}\beta_g + \boldsymbol{\beta}_w}{\sqrt{1+c^2\beta_g^2}} = \boldsymbol{\gamma}_w.$$

Then

$$\boldsymbol{\beta}_w = \boldsymbol{\gamma}_w \sqrt{1 + c^2\beta_g^2} - b\boldsymbol{\theta}\beta_g. \tag{B.7}$$

□

Observe that Equations B.2 and B.4 imply an upper bound on R_f^2, the predictive power of the future PGI. Let $R_{G_c,w}^2$ denote the R^2 of the regression of G_c on $\mathbf{W}$, and observe that $R_{G_c,w}^2 = \mathrm{Var}[\mathbf{W}\boldsymbol{\theta}]$ (since $\mathrm{Var}[G_c] = 1$). From this and the fact that $\sigma_V^2 \geq 0$, it follows that

$$\sigma_\epsilon^2 \leq 1 - \mathrm{Var}[\mathbf{W}\boldsymbol{\theta}] \implies 1 - \frac{R_c^2}{R_f^2} \leq 1 - R_{G_c,w}^2 \implies R_f^2 \leq \frac{R_c^2}{R_{G_c,w}^2}.$$

C Estimation of the single-disease model

Section 4.3 describes how we estimate the single-disease model for Scenarios 1 and 2. For Scenarios 3L and 3U, as also described there, we first estimate R_c^2, σ_ϵ^2, and $\mathbb{P}_W$. To estimate the remaining parameters $(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_V^2)$, we use maximum likelihood estimation. To simplify the analysis, write the vector of parameters as $(\sigma_T^2, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_V^2)$, where $\sigma_T^2 = \sigma_\epsilon^2 + \sigma_V^2$ and $\boldsymbol{\gamma}$ are the coefficients of the probit regression of L on G_c and $\mathbf{W}$ (Equation B.5). Written this way, $\boldsymbol{\beta}$ is a function of the other parameters, as given by Equations B.6 and B.7, and σ_V^2 is a function of σ_T^2.

We can decompose the likelihood function:

$$\begin{aligned} &f(L = l, G_c = g_c, \mathbf{W} = \mathbf{w} | \sigma_T^2, \boldsymbol{\gamma}, \boldsymbol{\theta}) \\ =&f_w(\mathbf{w}) \cdot f_{LM}(g_c | \mathbf{w}; \sigma_T^2, \boldsymbol{\gamma}, \boldsymbol{\theta}) \cdot f_{PM}(l | g_c, \mathbf{w}; \sigma_T^2, \boldsymbol{\gamma}, \boldsymbol{\theta}). \end{aligned}$$

The distribution of G_c conditional on $\mathbf{W} = \mathbf{w}$ is given by the linear model (LM) in Equation B.3 and the distribution of L conditional on $G_c = g_c$ and $\mathbf{W} = \mathbf{w}$ is given by the probit model (PM) in Equation B.5. Therefore, the likelihood can be simplified to

$$f_w(\mathbf{w}) \cdot f_{LM}(g_c | \mathbf{w}; \sigma_T^2, \boldsymbol{\theta}) \cdot f_{PM}(l | g_c, \mathbf{w}; \boldsymbol{\gamma}).$$

Therefore, the log-likelihood function is a constant plus

$$\log f_{LM}(g_c | \mathbf{w}; \sigma_T^2, \boldsymbol{\theta}) + \log f_{PM}(l | g_c, \mathbf{w}; \boldsymbol{\gamma}).$$

Thus, we can fit the model by maximum likelihood separately for $(\sigma_T^2, \boldsymbol{\theta})$ with a linear model and for $\boldsymbol{\gamma}$ with a probit model. $\boldsymbol{\beta}$ and σ_V^2 can then be calculated.

For inference, note that the log likelihood has no interaction terms between $(\sigma_T^2, \boldsymbol{\theta})$ and $\boldsymbol{\gamma}$. Therefore, the Fisher information matrix for these parameters is block diagonal, so $(\sigma_T^2, \boldsymbol{\theta})$ and $\boldsymbol{\gamma}$ are asymptotically normal and independent, with the standard errors from their separate estimation. Because $\boldsymbol{\beta}$ and σ_V^2 are smooth functions of the other parameters, they are also asymptotically normal. We estimate the full variance-covariance matrix for the parameters $(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_V^2)$ with the delta method. Because we are using maximum likelihood, the estimator is asymptotically efficient.